1-1-2004

An investigation of alternative approaches to scoring multiple response items on a certification exam.

Xiaoying Ma
University of Massachusetts Amherst

AN INVESTIGATION OF ALTERNATIVE APPROACHES TO SCORING MULTIPLE RESPONSE ITEMS ON A CERTIFICATION EXAM

A Dissertation Presented

by

XIAOYING MA

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirement for the degree of

DOCTOR OF EDUCATION

February 2004

School of Education

AN INVESTIGATION OF ALTERNATIVE APPROACHES TO SCORING MULTIPLE RESPONSE ITEMS ON A CERTIFICATION EXAM

A Dissertation Presented

by

Xiaoying Ma

Approved as to style and content by:

Stephen G. Sireci, Chair

Ronald K. Hambleton, Member

Beverly P. Woolf, Member

Andrew Effrat, Dean
School of Education

DEDICATION

To my husband, Li

whose love has sustained me throughout my graduate studies.

To my daughter, Ying Ying,

who makes me want to be a better person.

ACKNOWLEDGEMENTS

I wish to take this opportunity to acknowledge those individuals who are instrumental in my quest for a doctorate in education. First, I would like to thank the chair of my committee, Stephen G. Sireci, for his guidance and support throughout this research. Professor Sireci's passion for his own work and his critical approach to evaluating ideas make him an invaluable mentor; without his sound advice, this work would not have been possible. My sincere appreciation goes to Professor Hariharan Swaminathan, who, four years ago, opened the door to the wonderland of psychometrics for me. His encouragement and guidance helped transform me from a novice who knew little about testing and measurement into a confident researcher who has done many a project in the field of research and psychometric methods. I am also deeply indebted to Professor Ronald K. Hambleton, who never failed to offer valuable advice as well as words of encouragement at every point in my program of study. Apart from being a wonderful mentor, he also provided me with the opportunities to experience a wide range of projects that comprised a concrete part of my academic study; these experiences have made me a better researcher. I would also like to express my gratitude to Professor Beverly Park Woolf, for willing to serve on my committee and for her constructive advice on the dissertation.

I would like to thank all the students in the Research and Evaluation Methods Program who have helped me in various ways during the duration of my graduate study; I benefited tremendously from my interactions with these talented and wonderful fellow students. My steady progression through the program is due, in part, to each one of the

following friends and colleagues: Kristen Huff, Mike Jodoin, Saba Rizavi, Frederic Robin, Dehui Xing, their intellectual generosity has helped me overcome the great many obstacles encountered in my academic studies. My experience in REMP truly would not have been blissful, however, had I not the unique bond of friendship with Mary Zanetti, whose unconditional acceptance and intellectual camaraderie had made my four-year academic journey a pleasant and smooth sailing. Her presence as a fellow graduate student also helped create a friendly, supportive, and egalitarian atmosphere in REMP where intellectual discussions were facilitated and student interactions encouraged.

Last but not least, I would like to thank my husband, Li, for always being there, loving and supporting me. If an old saying holds the truth -- that searching for gold is merely a poor man's dream; only searching for love is a king's dream, then I am as rich as a queen with his love. I also want to thank my daughter Ying Ying, who has taught me about the beauty of innocence.

ABSTRACT

AN INVESTIGATION OF ALTERNATIVE APPROACHES TO SCORING MULTIPLE RESPONSE ITEMS ON A CERTIFICATION EXAM

FEBRUARY 2004

XIAOYING MA, M.A., NORTHWEST UNIVERSITY, CHINA

M.Ed., UNIVERSITY OF MASSACHUSETTS AMHERST

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Stephen G. Sireci

Multiple-response (MR) items are items that have more than one correct answer. This item type is often used in licensure and achievement tests to accommodate situations where identification of a single correct answer no longer suffices or where multiple steps are required in solving a problem. MR items can be scored either dichotomously or polytomously. Polytomous scoring of MR items often employs some type of option weighting to assign differential point values to each of the response options. Weights for each option are defined a priori by expert judgments or derived empirically from item analysis.

Studies examining the reliability and validity of differential option weighting methods have been based on classical test theory. Little or no research has been done to examine the usefulness of item response theory (IRT) models for deriving empirical weights, or to compare the effectiveness of different option weighting methods. The purposes of this study, therefore, were to investigate polytomous scoring methods for MR items and to evaluate the impacts different scoring methods may have on the reliability of the test scores, item and test information functions, as well as on

measurement efficiency and classification accuracy. Results from this study indicate that polytomous scoring of the MR items did not significantly increase the reliability of the test, nor did it increase the test information functions drastically, probably due to 2/3 of the items being multiple-choice items, scored the same way across comparisons. However, substantial increase in test information function at the lower end of the score scale was observed under polytomous scoring schema. With respect to classification accuracy, the results were inconsistent across different samples; therefore, further study is needed. In summary, findings from this study suggest that polytomous scoring of MR items has the potential to increase the efficiency (as shown in increase in test information functions) of measurement and the accuracy of classification. Realizing these advantages, however, will be contingent on the quality and quantity of the MR items on the test. Further research is needed to evaluate the quality of the MR items and its effect on the effectiveness of polytomous scoring.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Multiple-choice (MC) items are perhaps the most commonly used objective measure of knowledge, ability, or achievement in educational testing. A typical multiple-choice item consists of a stem and four to five response alternatives, only one of which is considered to be the "correct" or "keyed" option. MC items are preferred in many standardized tests with good reasons. First, MC items allow for broader content areas to be tested when compared to other formats such as open-ended questions and essays, hence, the use of MC items generally increases the validity of test score interpretations. Second, sufficient numbers of MC items, if properly constructed, generally produce high reliability of test scores. Last but not least, MC items can be easily pretested, stored, administered, and objectively scored. Thus, the use of MC items can reduce the costs associated with test development, administration, and scoring. Because of these advantages, MC items are the most popular selected-response items used by many testing programs.

Multiple-choice items are often scored conventionally with a value of 1 given to correct responses and a value of 0 for incorrect responses (including blank and omitted items), incomplete or partially correct responses are treated as wrong. This scoring scheme is often called number right scoring. Though appealingly simple, number right scoring of MC items has been criticized for encouraging examinee guessing on the tests. Moreover, number right scoring does not differentiate examinees with various levels of

(partial) knowledge (at the individual item level). Since guessing on the part of examinee contributes error variance to the observed scores and since failure to differentiate examinees on account of partial knowledge decreases measurement precision, the validity of the test scores and its use for decision-making purposes (e.g., selection, admission, certification, licensing, etc) are threatened. In attempts to reduce the scope of guessing and to extract partial information from examinee's responses to any given item, alternatives to the conventional number right scoring have been developed and their strengths and weaknesses studied for the past 50 years. However, before we get into a discussion on these methods, it is important to understand the effect guessing and partial knowledge can have on the validity of MC tests, as well as the significance for testing programs to address these effects during the examination and/or at the scoring stage.

It has long been recognized that examinees vary in their willingness to guess at answers on a test. Consider two examinees on a 100- item MC test: both know the answer to 60 items and are unsure of the answer to the other 40 items. One examinee leaves the 40 items unanswered, whereas the other guesses at the answers to the 40 items. If there are four choices per item, then the one that guesses is likely to get 10 of the 40 items correct by chance alone. If number right scoring were applied, one examinee would have a total score of 60, whereas another a score of 70. Thus, the two examinees would have different observed scores irrespective of their equivalency on the construct measured by the test (Crocker & Algina, 1986). Consequently, any inferences derived from these scores are likely to be biased due to the construct-irrelevant variance (guessing). In this case, the validity of test scores and its use is compromised.

Susceptibility to guessing is not the only problem that plagues conventional number right scoring of MC items. Equally problematic is its insensitivity to distinguish various levels of partial knowledge contained in examinee responses. Although some researchers adopt an all-or-none attitude, arguing that an examinee should receive a score if, and only if, the examinee has complete understanding of the test questions, most agree that partial knowledge should be taken into account when scoring examinees' responses to an item. Since knowledge is a not a dichotomous variable (Hutchinson, 1982), which the number right scoring apparently suggests, categorizing it into knowledge (which invariably leads to correct responses) and lack of knowledge (which leads to omission or excessive guessing) not only contradicts the psychological foundations on knowledge (Budescu & Bar-Hillel, 1993), empirical studies also proved the dichotomization to be unsound in real testing situations (Ben-Simon et al., 1997). Under conventional number right scoring rule, an examinee's responses to an item are characterized into knowledge or guessing categories, the continuity nature of knowledge is distorted, and intermediate levels of knowledge between the two extremes are ignored. As a result, examinees may receive identical item scores on a MC item regardless of the varying degrees of knowledge they may have about that item. Thus, like guessing, failure to assess partial knowledge also impairs the validity of the test score interpretations.

In view of the weaknesses associated with conventional number right scoring for MC items, efforts have been directed at finding alternative scoring algorithms that can resolve the problems discussed above. These alternative scoring strategies include those that attempt to discourage guessing on MC tests, such as formula scoring, and those that

assess different levels of partial knowledge, such as confidence weighting, answer-until-correct testing, option weighting, and elimination and inclusion scoring [see, for example, Ben-Simon et al. (1997) for a detailed review on these methods]. It should be noted, however, that some scoring strategies seek both to discourage examinees' guessing and to enhance the use of partial knowledge.

Modern test theory (i.e., item response theory) takes a different approach to addressing guessing and partial knowledge representation on MC tests. The three-parameter logistic IRT model explicitly incorporates a pseudo-guessing parameter to account for guessing by low ability examinee on MC tests. With respect to assessing partial knowledge on MC tests, Samejima's (1979) multiple-choice model, Bock's (1972) nominal response model, and the multiple-choice model of Thissen and Steinberg (1984), have all been used to score MC tests. Thissen and Steinberg's multiple-choice model also includes a category function (denoted as "don't know" category) to account for guessing by low ability examinees. As alternatives to the classical test theory models, these polytomous IRT models provide an attractive means for assessing examinees' varying levels of partial knowledge.

Scores of studies have been conducted to investigate the reliability and validity of alternative scoring methods for MC items within the framework of classical test theory (Frary, 1980; Crocker & Algina, 1986; Jaradata & Tollefson, 1988). Results from these studies are mixed at best, however. Some researchers report an increase in the reliability and validity coefficients of the tests when these methods were applied (Hanna, 1975; Wilcox, 1981), whereas others have found that they had not resulted in increases in reliability and validity; or even if there were increases in reliability and

validity, they were far from dramatic and were often offset by added cost of complex scoring (Raffeld, 1975; Jaradat & Tollefson, 1988).

Moreover, these alternative scoring schemes are often accompanied by complex instructions (e.g., eliminating a number of alternatives, assigning probabilities to response alternatives, and so forth) that are not always followed by examinees when responding on the MC tests. Hutchinson (1982) showed that even highly motivated examinees do not always understand, remember, or follow the instructions of the simplest scoring rule, and thus do not obtain full credit for their true level of knowledge. In addition, empirical studies suggest that examinees' specific personality traits have substantive influence on their performance in the confidence testing situations. In this case, instead of minimizing the effect of irrelevant variance, the correction rule may actually add new sources of measurement error.

In contrast to the substantive number of studies on polytomous scoring of MC tests within the framework of classical test theory, studies on IRT-based polytomous scoring methods are still evolving. While some researchers find that polytomous scoring yields considerably more IRT information than dichotomous scoring (Samejima 1976; Thissen, 1976; Donoghue, 1994), others raise questions about the usefulness of polytomous scoring methods. A study by Yamamoto and Kulick (1992) compared both dichotomously and polytomously scoring methods for the same items in terms of the relative information function and found the polytomously scored items contained, on average, slightly less information than did the dichotomously scored items.

General disappointment with alternative scoring rules for MC items has prompted test developers and researchers to search for alternative item formats rather

than special scoring rules. One such format is the multiple response format, which, because of its intrinsic features, has the potential to minimize guessing as well as to enhance partial knowledge representation on a test.

The multiple response (MR) item type is a variation on the typical MC format. A MR item consists of a stem and four or more alternatives (typically 6 is the maximum, but in some cases, the number of alternatives could be up to 20 or more); any number of which can be the keyed answers. Generally, two types of MR items are used in testing practice, in one of which the number of correct answers is specified to examinees in advance; the other more open-ended type requires examinees to mark all answers that are correct without mention of the exact number of correct answers. MR items have been given various names depicting the distinct feature of MR items—multiple correct answers. Among them are: multiple-correct-multiple-choice item, multiple-mark item, multiple-multiple-choice item, and key-feature item, to name but a few that have been used by measurement experts. MR items bear close resemblance to Type K and Multiple True False items, which are two other variants of MC items. It should be noted, however, that the MR format differs from the Type K and MTF (see chapter two for an in-depth discussion on the differences among these three formats) format. MR items are preferred over these two formats by many testing programs.

MR items can effectively reduce guessing by expanding the range of possible response options. The chance of guessing on an item is determined by the combination formula:

$$\binom{n}{r}\frac{n!}{r!(n-r)!}$$

where n is the number of response options and r is the options selected from n without replacement.

Consider a four-option item, if there is only one correct answer, an examinee has a 25% chance of answering it correctly; however, if there are two correct answers, the chance that an examinee answers it correctly by random guessing is reduced to one sixth. Thus, the scope of guessing in the MR format is reduced to a lesser degree.

More importantly, MR items offer flexibility in test construction that can accommodate situations when more than one alternative can be the keyed answer, or when multiple steps are required in solving a problem. Under these circumstances, identifications of single correct responses no longer suffice. MR items, by allowing examinees to answer at different levels of sophistication, enable a more meaningful observation of different stages of cognitive process in a systematic manner, thus providing a more valid measurement of examinees' true state of knowledge.

There is a long history of theoretical and empirical work that forecasted the viability of MR items. For instance, studies by Cronbach (1941), Dressel and Schmid (1953) on the MR format, and studies by Coombs (1953), Coombs, Milholland, and Womer (1956) on elimination scoring (an alternative response and scoring method for MC tests; examinees are instructed to eliminate all distractors that they can identify as incorrect) suggest an advantage to the MR framework in that it provides a means for examinees to make optimal use of their partial knowledge in responding, thus allowing a more meaningful observation of examinees' varying degrees of partial knowledge and affording a finer discrimination among them. While most of the early studies on MR items were carried out with small-scale classroom tests, they helped establish a

theoretical foundation for the subsequent adaptation of MR items in large-scale high-stakes testing conditions.

In recent years, MR items have been used in large-scale state assessment programs as an alternative to the more costly performance assessment tasks. One such application is the multiple-mark items used in Kansas Reading and Mathematics Assessments. Studies bearing on the reliability and validity of the MR format as applied in large-scale state assessment program have been reported (Glasnapp & Poggio, 1994; Pomplun & Omar, 1997) and are discussed later.

Perhaps the most frequent use of the MR format as an objective testing tool is found in the health professions. Known as "key-feature" (Page, Bordage, & Allen, 1995) problems in medical testing, MR items are considered to be an effective measure of clinical problem-solving skills, as clinical problems, which often have multiple solutions, are not amendable to the standard MC types of assessment.

It should be noted that most MR items are nested with a testlet -- a measurement unit containing a number of items related to a single content area (Wainer & Kiely, 1987), as a testlet can provide a coherent measure of complex problem-solving skills and an explicit framework for awarding partial credit. Examples of the testlet-based MR items include the key-feature problems mentioned above and the multiple-mark items used in Kansas Reading and Mathematics Assessments.

The recent movement from traditional paper-and-pencil testing to computer-based testing also generates considerable interest in innovative item types, MR format being one of them (Parshall, Davey, & Pashley, 2001). Uses of MR format in computer-based testing have been reported in the literature. O'Neill and Folk (1996) reported the

use of MR ("select two of the following alternatives") in a computerized test, Parshall, Stewart, & Ritter (1996) reported the use of the more open-ended MR item ("click on all correct responses"), and Jodoin (2001) reported the use of both types of MR items ("select 2 of 5 alternatives" and "choose all that apply") in the Microsoft Certification Program. As many testing programs have implemented computer-based testing or are in the process of doing so, it is expected that innovative item formats like the MR item type will attract more attention and be used more often than now in operational testing.

From the foregoing discussion it is clear that MR format is an attractive alternative to the standard MC format. The fact that it has not seen widespread use like the MC items in educational testing is due, perhaps, to the technical difficulties involved in constructing appropriate MR items and scoring them.

1.2 Statement of the Problem and Its Significance

MR items can be scored in several different ways. Scoring on an all-or-none basis (one point if all the correct answers and none of the distractors are selected, and zero points otherwise), or scoring each alternative independently (one point for each correct answer chosen) are commonly used methods. Both methods, however, have disadvantages. With the first method, a student who correctly identifies all but one of the answers receives the same score as a student who cannot identify any of the answers.

The second method dichotomizes the MR item, with each response alternative virtually becoming an individual item and scored using the conventional number right scoring rule. Although this scoring procedure can be considered as a polytomous

scoring rule (at the response option level), representing a step forward from the dichotomous scoring in that scores are more representative of each student's achievement (at the option level), it fails, nevertheless, to achieve the desired measurement precision intended by MR format, mainly due to the fact that treating alternatives as independent items will overestimate reliability because of the dependency among the response alternatives (Sireci, Thissen, & Wainer, 1991).

Since the driving principle that initiated the use of MR format in testing is its effectiveness in differentiating between examinees with different levels of knowledge, the appropriate scoring rule for MR items should, consequently, be of a polytomous nature (i.e., to distinguish between examinees with different levels of partial knowledge), and be accurate (i.e., to avoid the local dependency problem). The underlying assumption for polytomous scoring of MR items is that a systematic relationship exists between distractors and correct answers (Levine & Drasgow, 1983) and scoring algorithms should make use of this differential information represented in different response alternatives to improve the precision of measurement. Although pertaining to MC test, this assumption is applicable to MR scoring as well for these two formats share a lot of common characteristics.

Both classical test theory and item response theory provide ways to score test items polytomously. When items are scored polytomously under a classical test theory model, some form of option weighting occurs, i.e., each response alternative of an item is assigned a differential point value to reflect its degree of correctness to the keyed answer. The point value assigned thereby is called the weight for that option and the process is referred to as option weighting. An examinee's score on an item depends on

which response alternatives he/she chooses. Consider a four-alternative MR item with two correct answers (the number of correct answers may or may not be specified in advance), alternatives A and D, for example, an examinee who selects both A and D will receive the maximum score point for this item, whereas another examinee who selects only A may receive a fraction of the maximum score points. Still, a third examinee who selects A and a distractor B may receive a fraction of the maximum score points, too; though his/her score on this item will differ from that of the second examinee because differential scoring weights are placed on the response alternatives they chose. Since the weighted scores are more representative of the examinee's true level of knowledge, the validity of test score interpretations is enhanced. In practice, differential option weighting is achieved by a judgmental procedure or through empirical item analysis.

When weights are determined a priori by judges or based on a theory of the structure of knowledge (e.g., Smith, 1987), they are typically called a priori or logical weights. A priori weights can be determined by simply averaging judges' ratings for each option (Downey, 1979) or scaling judges' rankings to obtain weights (Patnaik & Traub, 1973). Logical weighting of response options has been reported to increase reliability slightly (Hambleton, Roberts, & Traub, 1970; Patnaik & Traub, 1973). The results for validity vary: some found an increase in predictive validity (Hambleton et al., 1970), whereas others found no meaningful difference between option-weighted scores and number right scores (Downey, 1979).

Option weighting based on the responses of examinees themselves is called empirical option weighting. Empirical option weighting generally falls into two

categories: linear option weighting and nonlinear option weighting. Linear option weighting procedures are developed within the framework of classical test theory, which involves using linear methods to obtain scoring weights for response options. These weights are computed on the basis of the option's attractiveness, average standardized score of examinees selecting an option, the point-biserial correlations between choosing each option and total score, as well as other similar procedures (Guttman, 1941; Davis & Fifer, 1959; Serlin & Kaiser, 1978). These weighting procedures typically seek to maximize the internal consistency reliability of the test.

A major disadvantage of these linear option-weighting schemes is that the weights derived thereby are sample dependent; it is crucial, therefore, to cross-validate the weights obtained from one sample against that from another sample. A better solution to this problem, perhaps, is to use item response theory models to weight response options differentially. This school of option weighting procedures is based on polytomous IRT theories and models. In contrast to dichotomous IRT models, which model the probability of correct response, polytomous IRT models model the probability of selecting each response category, thereby information contained in incorrect responses also contribute to the scoring function, enabling a more precise estimates of examinee ability (Baker, 1992).

Several polytomous IRT models, including Bock's nominal response model, Samijema's multiple-choice model, and the multiple-choice models of Thissen and Steinberg, have been used to study the differential functioning of response alternatives. Findings from empirical studies suggest that option weighting increased accuracy of ability estimation over the lower half of the score range.

From the discussions on classical and IRT-based polytomous scoring of MC items in the above, a general theme emerged: Most of the studies found an increase in the internal consistency reliability. Results on validity varied: some studies found an increase in validity, whereas others reported a decrease as a result of applying option weighting to MC tests.

Empirical findings notwithstanding, the theoretical appeal of polytomous scoring methods (e.g., minimizing construct-irrelevant error such as guessing, enhancing partial knowledge representation) upholds its usefulness in scoring MC tests. The extent to which it can be generalized to score MR items remains to be seen as no study of this sort has been reported in the literature.

To date, most operational MR tests are scored dichotomously either at the item level (i.e., score each item as right or wrong) or at the option level (i.e., score each response option as right or wrong). Both are considered as being inadequate to fully assessing partial knowledge. For the few testing programs that utilize polytomous scoring procedures, option weights are determined a priori by judges and are not subjected to empirical analysis. Although expert and consensus judgment is critical in determining the rightness and wrongness of options, there is no way to know whether it is the best method in all situations. Option analysis is important in this regard as it can detect potential errors of judgment and uncover the inadequate performance of distractors (Haladyna, 1997).

The present study represents an effort to explore alternative scoring methods for MR items and compare different option weighting methods with respect to measurement efficiency and classification accuracy in the context of a certification test.

The idea here is that by studying the performance of different polytomous scoring methods, we may be able to find a viable and more efficient alternative approach to scoring MR items empirically, should the judgmental weighting becomes impractical or unreliable. It is also hoped that this research will help us understand the practicality of IRT approaches to scoring MR items. It is on these aspects that the importance of this study lies.

1.3 Purpose of the Study

The purposes of the study, therefore, are: (1) to investigate alternative approaches to scoring MR items, (2) to compare different polytomous scoring methods in terms of measurement accuracy and efficiency, and classification accuracy, and 3) to examine the congruence of the option weights obtained from different option analyses. The hypotheses of the study are: 1) polytomous scoring is expected to improve the precision of measurement with respect to the psychometric properties of the test, 2) option weights obtained from different weighting methods will be highly correlated, and 3) polytomous scoring is expected to improve the accuracy of classification.

CHAPTER 2

REVIEW OF THE LITERATURE

2.1 Overview of Multiple Response Items

Multiple response (MR) items are an extension to the usual one-answer and one-response (one keyed alternative) multiple-choice (MC) items by having more than one correct answer per item and requiring examinees to select all the correct answers. This type of item consists of a stem and several response alternatives (typically the maximum number is six, but in some cases it could be up to 20 or more), any number of which could be the keyed alternatives. Various names have been given to this type of items, including "multiple-multiple-choice" (Cronbach, 1941), "multiple-answer" (Dressel & Schmid, 1953), and "multiple- mark" (Pomplun & Omar, 1997) items. Two types of MR item are commonly used in testing practice, the first has the number of correct answers per item specified to examinees in advance; the other more open-ended type requires examinees to mark all correct answers without mention of the exact number of correct answers. These two types of MR items may result in cognitively different tasks for examinees, though, to date, no study has been undertaken to compare these two item types. MR items have been used to accommodate situations when more than one answer could be correct or when multiple steps are required to solve a problem.

2.1.1 Comparisons of Three Multiple Response Formats

MR items share a common characteristic with the multiple True-false (MTF) and the Type-K items: that is, all three types of items have more than one correct answer per item. However, they differ in item representation and response mode. Since all three multiple response formats have been used in various tests, it is necessary, therefore, to discuss the characteristics of each of the multiple response formats and to compare them with respect to their psychometric properties (reliability and validity). Figure 2.1 depicts the characteristics of MR, Type K, and MTF formats.

A Type K item consists of a stem, a list of potentially correct answers referred to as the primary responses, and a list of combinations of the primary responses, such as "I, II only," "I, III only," "All of the above," "None of the above," referred to as the secondary choices. Examinees are instructed to select from the secondary choices, permitting only one mark on the answer sheet as the MC item, thus Type K items can be scored as MC items, regardless of its multiple-correct-answer feature.

The Type K item was originally introduced by Educational Testing Service, and, later, was adopted for use in medical testing by the National Board of Medical Examiners (NBME, Hubbard, 1978), which designated it as the Type K item. It has also been referred to as complex multiple-choice item (CMC) in the literature. An inherent flaw of the Type K format is that it provides clues to examinees, i.e., knowing that one option is absolutely correct or incorrect helps the examinee identify the correct option by eliminating distractors. Cluing introduces error variance into test scores and consequently reduces the validity of the test (Albanese, 1993; Haladyna & Downing, 1989). To illustrate this point, consider the sample Type K item in Figure 1, knowing

that primary response C is incorrect will lead an examinee to eliminate second choices 2, 4, and 5. As a result, the examinee's chance of identifying the correct answer increases to 50%.

Comparisons of the reliability and validity of Type K items with MR items have shown that Type K items tend to have lower reliabilities and validities than MR items (Albanese, 1993). In addition, this format takes up more space on the page and requires more reading time. Thus, the number of items that might be included in a test is limited and the sampling of content is negatively affected (Haladyna & Downing, 1989). Because of these deficiencies, many testing programs such as the National Board of Medical Examiners decided to discontinue the use of such items (Albanese, 1993).

A better replacement for the Type K item is a multiple true-false item. A multiple true-false item consists of a stem that is an incomplete statement, followed by four or five response alternatives that independently complete the stem (Hubbard, 1978; Frisbie, 1992). The examinee is instructed to respond to each of the response alternatives as TRUE or FALSE. The stem can also be in a question format such as "Which of the following is true (or false)? (Ebel, 1978)" MTF items have been called the Type X item in medical testing (Hubbard, 1978) to distinguish it from the Type K item.

Research on MTF items can be traced back to 1930, though its evolution as an effective testing format only gained popularity in the 1980s, fueled in part by the need to find a suitable alternative to Type K items (Frisbie, 1992).

I. MR Item

Which of the following components are hardware storage devices used on PCs (choose three)?

*(a) Diskette

(b) Central Processing Unit (CPU)

(c) Databases

*(d) CD-ROM

*(e) Hard disk

II. Type K Item

Which of the following components are hardware storage devices used on PCs?

Primary Responses

(I) Diskette

(II) Central Processing Unit (CPU)

(III) Databases

(IV) CD-ROM

(V) Hard disk

Secondary Choices

(1) I, II only

(2) I, III, IV only

*(3) I, IV, V only

(4) II, III, V only

(5) All of the above

III. MTF Item

The hardware storage device used on PCs is:

(a)	Diskette	T*	F
(b)	Central Processing Unit (CPU)	T	F
(c)	Databases	T	F
(d)	CD-ROM	T*	F
(e)	Hard disk	T*	F

Figure 2.1 Sample MR, Type K, and MTF Items

Frisbie summarized studies on MTF items and concluded that: (1) MTF items tend to yield more reliable scores than MC items; (2) MTF items measure the same underlying construct as content-parallel MC items; (3) MTF items tend to be more difficult than MC items (the order of difficulty (from hardest to easiest) is MR, CMC, MTF, and MC); and (4) Examinees generally preferred MTF format to MC and CMC formats. Based on the findings, Frisbie recommended that MTF items be studied further.

MR items bear close resemblance to MTF items. In fact, the more open-ended MR item (mark all correct answers) is virtually identical to MTF in that both require examinees to make a true/false judgment for each of the response alternatives, the only difference being that under MR format, examinees do not need to mark the distractors as FALSE. Still, the MR items are preferred over MTF items for several reasons. First, it is believed that MR items may engage examinees in a more involved and extended thought process, " A student may be forced not only to see the relationships existing between a stem and the responses, but also to reconstruct his thinking as he looks at each response in relationship to the other responses of the item" (Dressel & Schmid, 1953, p. 581). Dressel and Schmid postulated that this complex thought process might not occur with the true-false testing.

Second, there are known response biases in the MTF format. Known as the tendency "acquiescent," which suggests that when in doubt, an examinee is likely to respond "true" rather than "false" to an answer to a MTF item (Cronbach, 1941), these biases result in lower reliabilities and validity coefficients for scores from "true" items (Cronbach, 1941; Grosse & Wright 1985). However, it should be pointed out that the

use of MR format is likely to introduce response biases, too, though the bias may be of a different nature. It has been observed that under MR format, examinees disproportionately choose not to mark when in doubt, resulting in lower reliabilities and validity coefficients for scores from incorrect options (Glasnapp & Poggio, 1994; Pomplun & Omar, 1997).

One drawback of the MR format is that omission and "false" are confounded blank responses (i.e., whether a blank response is an omission or a judgment of false is unknown). Despite this disadvantage, however, the advantages discussed above lead to the recommended use of the MR format in educational testing. The next section offers an extensive review of the use of MR items in various tests.

2.1.2 Use of Multiple Response Items in the Literature

Use of the MR format was suggested as an effective mode of test construction by Orleans and Sealy (1928, pp. 223-226), who designated this type of item as multiple-choice plural-response question, to distinguish it from the single-response multiple-choice type of question. Orleans and Sealy demonstrated that MR items could be used to measure from rote knowledge such as the recall of information using discrete MR items, to more complex cognitive processes such as reasoning using MR items that are connected to common stimuli. The difficulties in scoring MR items were also discussed by Orleans and Sealy.

Several studies contributed theoretically and empirically to the line of research on MR items. For instance, the studies of Cronbach (1941) and Dressel and Schmid (1953) suggest an advantage of the MR framework in that it provides a means for

examinees to make optimal use of their partial knowledge in responding, thus enhancing the validity of the test score interpretations. Although these studies were carried out with small sample classroom tests, they provide a theoretical ground on which subsequent adaptations of MR items in large-scale testing are based. Because of their importance in the literature, they are well worth a brief discussion here.

In an experimental study, Cronbach (1941) compared the MR format with MTF format. Both types of items were presented in a MTF like format including a stem, which is an incomplete statement, and five response alternatives, which are to independently complete the stem. Twenty-two items of each type were administered to about 60 students. A major difference between the two formats is that examinees were not required to mark the distractor as false in the MR format. The number of correct answers to a MR item varied from item to item (from one to several, and in some cases, none) and it was left to the examinees to make that judgment. Because students were instructed to guess when uncertain, omissions were not a problem. Cronbach found no statistically significant difference between the two formats in terms of the time required for completion of the test (in minutes, MR = 30.8 vs. MTF = 31.2), the difficulty of tasks (mean scores, MR = 26.9 vs. MTF = 26.3), and reliability (MR = 0.53 vs. MTF = 0.428) and validity (MR = 0.62 vs. MTF = 0.598). The reliability of MR format was slightly higher than that of MTF format, however. Since the tendency of examinees to mark "true" rather than "false" when in doubt was observed under the MTF testing, Cronbach recommended the use of a MR format over a MTF format for slightly higher reliability.

Dressel and Schmid (1953) empirically investigated several variants of the MC format, one of which is the MR format. Two types of MR items were studied in their experiment. The first, referred to as the two-answer item, consisted of a stem and five response alternatives; examinees were informed that the number of correct answers per item was two, hence the name two-answer item. The second type of MR items, referred to as the multiple-answer items, also consisted of a stem and five response alternatives; here, however, examinees were instructed that any number of which might be correct and they were to mark all the correct answers. Forty-four items of each type were administered to approximately 90 students. They concluded that MR items had a slightly higher reliability (0.78 and 0.76 for multiple-answer and two-answer items, respectively) than the MC format (0.70) and other MC variants (0.67 and 0.73 for free-choice and degree of certainty items, respectively). Moreover, Dressel and Schmid suggested that the multiple response type of items has the potential to measure various levels of partial knowledge and to afford a finer discrimination among examinees.

Since Cronbach (1941) and Dressel and Schmid's (1953) seminal work on MR items, research on this type of item and its usefulness in testing has been scarce, presumably because of the dominance of the MC format and the difficulties involved in constructing good MR items and scoring them. Few researchers have reported the use of MR items in their studies, compared to a myriad of studies on constructing MC items and using them in various tests. This undesirable state of affairs, however, was ended in the 1980s as a consequence of the educational reform, which, with its call for a more authentic and valid assessment of student learning outcomes, generated considerable interest in using MR items as an alternative to performance assessment tasks, as the MR

format can retain many of the beneficial aspects of the MC format and, at the same, measure higher order thinking skills at a lower cost than the performance assessment tasks. Applications of the MR items in large-scale state assessment programs have been reported in the literature. For instance, the Kansas State Assessment Program (Pomplun & Omar, 1997) uses the more open-ended MR item on reading tests at grades 3, 7, and 10, and on math tests at grades 4, 7, and 10. Approximately 30,000 students took each test at each grade level. Pomplun and Omar (1997) investigated the psychometric properties of the multiple-mark items and concluded that there is adequate reliability (for example, 0.73 to 0.78 when scored at the option level for reading tests) and validity evidence to support the use of MR format, and, because of its desirable features (e.g., allowing multiple correct answers, ease of scoring), it is a promising item format for use in state assessment programs.

It should be noted that the MR items used in these two assessment programs are nested within a testlet -- a measurement unit containing a number of items related to a single context or content area (Wainer & Kiely, 1987). The rationale is that only the testlet format is broad enough to provide a coherent measure of complex problem-solving skills and an explicit framework for measuring different levels of partial knowledge.

Perhaps one area in which the MR item format is frequently used is in medical credentialing. Page, Bordage, and Allen (1995) described the development of "key-feature" problems for use in the Canadian Qualifying Examination in Medicine, a licensing exam taken by all graduates of Canadian and foreign medical schools before practicing medicine anywhere in Canada except Quebec. According to Page et al., key

feature is defined as a critical step in the resolution of a clinical problem, and a key-feature problem consists of a clinical case scenario followed by questions that focus on only those critical steps. The appropriate responses to a key-feature problem could be one or several. One of the formats used for key-feature problems is the MR format in which a list of response options is presented to the examinees. Page et al. evaluated the psychometric properties of the key-feature problems and concluded that they are valid and reliable measures of clinical problem-solving skills and worthy of consideration by medical testing professionals. According to Page et al., the American College of Physicians and other medical schools have subsequently adopted the key-feature problems for use in their testing programs.

The recent movement from traditional paper-and-pencil testing to computer-based testing also generated interest in innovative item types that are adaptable to computerized test administration to take advantage of graphics and timing capabilities available through computer technology. One of the innovative item types is the MR format (Parshall, Davey, & Pashley, 2001). There are reports about the use of the two types of MR items in computerized testing (Jodoin, 2001; O'Neil & Folk, 1996). In 1996, Parshall et al. undertook a study to investigate the feasibility of using innovations such as graphics, sound, and alternative response modes in computerized tests. One section of their study was devoted to the evaluation of the MR format. The MR items used therein were the more open-ended ones ("select all that apply") and were scored dichotomously. Parshall et al. concluded that the psychometric functioning of the various item types appeared adequate and that examinees were largely positive about the computer examination. They suggest that future research on MR format should

focus on the effects of guided instruction (e.g., "select the best 3") and of partial-credit scoring. As many testing agencies have implemented computer-based testing programs or are in the process of doing so, it is expected that innovative item formats such as the MR format, will attract more scholarly attention and be used more often than now in operational testing.

From classroom tests to large-scale standardized tests, from paper-and-pencil testing to computer-based testing, the MR format is evolving as an attractive testing format. As MR items have been largely used in licensure, certification, and achievement tests where important decisions (e.g., placement, graduation, employment) are made, it is imperative that test scores be reliable, valid, and representative of examinees' true state of knowledge. The next section describes how MR items are scored in practice by different testing programs. The strengths and weaknesses of these scoring techniques are also discussed.

2.2 Existing Techniques for Scoring Multiple Response Items

There exist two classes of scoring methods for MR items, the first treats a MR item as an intact entity and utilizes the dichotomous scoring method. The second class consists of scoring techniques that treat each response option of a MR item as a separate entity and apply various formulas to correct for guessing. Table 2.1 summarizes the scoring formulae that have been used in practice.

Because of its intrinsic affinity to the standard MC item, MR items can be scored dichotomously like MC items to maintain the consistency of scoring between the two formats. Formula 1 applies the standard MC scoring rule to score MR items

Table 2.1 A Comparison of Existing Scoring Techniques

	Scoring Technique	Formula	A Sample Item Maximum Score: 5 Options: A, B, C, D, E Correct Answers: A, D, E An examinee selected: A, C, D	Application
1	All-or-none Scoring: Each item is treated as an intact entity)	Full score points if all correct options and none of the incorrect options are marked, 0 otherwise	0	
2	Formula Scoring: Each option is treated as an individual item	# of correct options marked + # of incorrect options unmarked	2 + 1	Kansas Assessment Program (Glasnapp & Poggio, 1994)
3	Formula Scoring: Each option is treated as an individual item	# of correct options marked / # of correct options in item, 0 if some options (e.g., dangerous actions) are marked	2/3	Key-Feature Problem (Page et al., 1995)
4	Formula Scoring: Each option is treated as an individual item	# of correct options marked – # of incorrect options marked (as correct)	2-1	Cronbach (1941) Dressel & Schmid (1953)

Note. Examination directions: Mark all the correct answers to the question.

dichotomously. Dichotomous scoring of MR items is on an all-or-none basis, i.e., full points are given if all the response options in the set are answered correctly (i.e., correct answers and none of the incorrect answers are selected for an item), and zero points otherwise (including blanks and partially correct answers). The rationale is that an examinee should receive score if, and only if, he has complete understanding of that item. Though under this scoring rule, the scope of guessing is largely reduced owing to

the special feature of the MR format (see CHAPTER 1 for a discussion on this), it is deficient, nonetheless, because it fails to reward partial knowledge. For example, a student who correctly identifies all but one of the correct answers receives the same score as a student who cannot identify any of the correct answers. Since MR items are designed with the intention to accommodate multiple correct answers and to allow examinees to respond at different levels of sophistication, this all-or-none scoring rule is counterintuitive, and, therefore, not commonly used by testing agencies.

Formulae 2, 3, and 4 belong to the second class of scoring methods that score each response option as an item independently of other options in the item set. All three formulas give partial credit if some of the options are correctly marked; however, they differ in how the selection of incorrect options is penalized. Since each response option is scored as right or wrong, an examinee has a 50% probability of answering it correctly by chance alone. As a result, the validity of the item score is likely to be severely reduced if guessing on the part of examinees were not corrected. The three scoring techniques take different approaches to addressing guessing and their impact on the possible score for a hypothetical MR item is shown in Table 2.1. Under Formula 2, one point is given for a correct response – that is, identifying a correct option as correct therefore marking it, and identifying an incorrect option as incorrect and not marking it; no penalty is given for an incorrect choice (it might be argued that there is a penalty through "opportunity loss" to gain potential points from the incorrect options), examinees typically obtain higher scores under formula 2. The Kansas Assessment Program (Glasnapp & Poggio, 1994; Pomplun & Omar, 1997) applies this formula in scoring MR items on the reading and mathematics tests.

Under Formula 3, a fraction (Because the maximum score for a key-feature question is "1", fractions instead of numbers are used in calculating item scores) of the maximum score point is awarded for partially correct answers and an examinee who selects 2 of 3 correct answers would receive a score of 2/3 for that question. However, a severe penalty (i.e., a score of 0) is exacted if some response options are chosen (Page, Bordage, & Allen, 1995). The principle is that the actions implied by these responses could cause grave consequences (e.g., life-threatening) hence no points should be given to an examinee who select these response options no matter what other responses this examinee makes. Under Formula 4, each correct option is weighted equally. Alternatively, according to Page et al., each correct option can be weighted differentially (e.g., larger weights can be assigned to some responses that are considered to be more important than others), though no discussion is provided as how to accomplish the task.

Of the three formulas discussed here, Formula 4 is the most stringent scoring rule. Formula 4 corrects for guessing by subtracting the number of incorrect choices from the number of correct choices. An issue regarding the use of Formula 4 is that under this formula, the possible score for an item could be negative. Consider, for example, the sample MR item in Table 2.1, if an examinee marks A, B, and C (correct answers are A, D, and E), he would receive a score of –1, thus formula 4 overcorrects for guessing and/or misinformation. Cronbach (1941), and Dressel and Schmid (1953) pioneered Formula 4 in their experimental studies.

The second class of scoring techniques is clearly superior to the all-or-none scoring rule, because these scoring methods allow examinees to make a judgment on

every option independently of other options, thereby providing a means for them to demonstrate their true level of (partial) knowledge; an examinee who correctly identifies two correct answers would receive more credit than the examinee who does not recognize any of the correct answers. Therefore, scores produced by these scoring procedures are more representative of each student's achievement, can afford finer discrimination between examinee who knows some of the correct answers, and the examinee who knows none of them.

The performance of formula scoring methods for MR items was evaluated by Hsu, Moss, & Khampalikit (1984). Six scoring formulas for MR items were compared in terms of difficulty, discrimination, reliability, and efficiency, using data from a college entrance examination (the College Entrance Examination of Taiwan). These six formulas vary in terms of the assignment of partial credit and the correction for guessing. Hsu et al. found that giving partial credit resulted in a slight increase in the reliability. With respect to correcting for guessing, the study found that formulas without correction performed at least as well as the formulas with correction.

Hsu et al.'s study lends support to the contention that MR items should be scored with a partial credit algorithm to better reward partial knowledge. However, the study suggests a disadvantage to the correction-for-guessing formulas because the value of formula scoring in producing more reliable and valid scores is dubious and the minimal gain that is realized is largely offset by the cost of complex scoring. Furthermore, correction-for-guessing formulas are often accompanied by instructions that interact with examinee's personality (e.g., propensity for guessing) and response strategies and may introduce errors that are unrelated to the construct being measured.

Therefore, researchers recommended against the use of formula scoring (Budescu & Bar-Hillel, 1993; Diamond & Evans, 1973).

Apart from the disadvantage discussed above, a more serious issue regarding scoring each option independently concerns the dependency among response alternatives. Since in the MR format the four or six response options are related to one item stem, there might be statistical dependence among these options. Put differently, these options are not locally independent, i.e., they share something in common even after eliminating the influence of the general common factor (e.g., ability, proficiency) from every item. Consequently, a fundamental assumption – the local independence assumption -- required by measurement modeling, both in classical test theory and item responses theory frameworks (Hambleton, Swaminathan, & Rogers, 1991; Yen, 1984), is violated in a test composed of locally dependent items.

When item alternatives are dependent, a score based on the separate alternatives will not contain the same amount of information as a score based on alternatives summed and then calibrated at the item level (Yen, 1993). As a result, methods to calculate reliability that treat the alternatives as independent items will overestimate reliability (Sireci, Thissen, & Wainer, 1991). Investigations of the local dependency problem with MR (and MTF format for this matter) have been conducted; the findings from empirical studies attest to the existence of local dependency among alternatives of a MR item (Wang & Acherman, 1994; Glasnapp & Poggio, 1994; Pomplun & Omar, 1997).

In view of the problems associated with existing scoring techniques for MR items, it is deemed reasonable to explore alternative scoring methods that can avoid

these problems while realizing the benefits intended by use of the MR format. Such methods should have the desired property of assessing different levels of knowledge while obviating the potential problem of local dependency resulting from scoring each option independently. Polytomous scoring algorithms that can encompass both the goals can be found within the framework of classical test theory and as well as in item response theory.

Classical test theory and item response theory (IRT) offer different ways of scoring an item polytomously. Under classical measurement theory, polytomous scoring is achieved through weighting response options of the item differentially such that a "more correct" response would be weighted more heavily than a "less correct" answer. Differential weighting of options is often scored-based, the weights being derived to maximize the internal consistency reliability of the test scores. Under item response theory, polytomous scoring is information-based, achieved by using polytomous IRT models to extract maximal information from item responses, including the amount of information conveyed by incorrect responses. A myriad of scoring methods have been developed in the past 50 years to score MC items polytomously, an in-depth discussion of which is presented in the following sections. While these methods deal mainly with MC items, it is possible to adapt them for use in scoring MR items.

2.3 Polytomous Scoring: Option Weighting

Classical polytomous scoring methods typically involve assigning differential point values to response options to reflect the relative correctness of each of the response alternatives. The point value that associates with each response alternative is

called the option weight for that alternative and the process is referred to as option weighting. Option weighting is based on the notion that reliability and validity should increase as a result of the finer discrimination among levels of ability afforded by differential option weighting procedures. It should be noted that option weighting differs from item weighting in that under item weighting each item may assume different point value, whereas under option weighting each item typically has a uniform point value. It has been recognized that there is no advantage to item weighting when a test contains more than 10 items that correlate positively with each other (Stanley & Wang, 1970), thus item weighting is rarely used in practice.

A variety of weighting schemes exist in the literature, which can be classified into two categories: objective weighting and subjective weighting. Objective weighting assigns weights based on examinees' responses to test items; subjective weighting requires examinees or experts to supply the weights for response options by confidence weighting, probability weighting, and logical weighting. Research related to these weighting schemes is presented in the following sections.

2.3.1 Subjective Weighting

Subjective weighting includes confidence weighting and probability weighting, both require examinees to assign weights to response options according to their belief in the correctness of these options. The difference is that the former requires examinees to select only the options they believe to be correct and indicate their degree of confidence about that, whereas the latter requires examinees to assign weights to each option according to the probability that it is correct. Confidence weighting can be regarded as a

simplified version of the probability weighting since with confidence weighting examinees only need to evaluate those options that are most likely to be correct (i.e., according to their belief). Under subjective weighting schemes, examinees choosing the same response may receive different scores for that item because of their indications of their degrees of confidence in their responses. It was proposed that that the reliability and validity of tests might be increased if the examinee assigns weights to the options according to his confidence in the correctness of each option (DeFinetti, 1965; Shuford, Albert & Massengill, 1966). However, results from empirical studies were disappointing (see, for example, Echternacht 1972, for a review). Questions regarding the validity of the subjective weighting methods were also raised. Research has shown that confidence is a personality trait that functions independently of other stimuli, and that weighting on this basis could result in an increase in measurement error variation (Ebel, 1965; Echternacht, 1972). Moreover, confidence and probability weighting involve complex response and scoring techniques, which, combined with other factors, diminishes the attractiveness of confidence weighting.

In recent years, however, several studies reevaluated confidence weighting and suggested that confidence weighting testing, if properly administered, could be a viable means for assessing partial knowledge (Ben-Simon et al., 1997; Holmes, 2002), but since the primary objective of the current research is to evaluate empirical option weighting methods, new developments on confidence and probability weighting are not germane to this review. For an in-depth discussion of these studies, the reader is referred to Ben-Simon et al (1997) and Holmes (2002).

2.3.2 Logical Weighting

Logical weighting refers to assigning weights to response options on a logical basis according to some prior belief about the correctness of the options. Hence, it is also called a priori weighting in the literature. Although not generally thought of as being a priori, both number right and formula scoring can be viewed as a priori weighting systems because equal weights are given to all response options.

There are several ways to derive logical weights for response options. Logical weights may be determined by a panel of judges based on the correctness of options or subsets of options to the keyed alternative. For example, some researchers simply instructed judges to rate each option on a 1-to-7 scale, and the judges' average rating was used as the weight for each option (Davis & Fifer, 1959; Downey, 1979). Others augment judges' ratings with scaling techniques, such as Thurstone's method of paired comparisons, or other multiple regression techniques such as facet analysis to ensure the quality of the derived weights (Jacobs & Vanderventer, 1968; Hambleton et al., 1970; Patnaik & Traub, 1973).

Another way of weighting options a priori is to instruct item writers to construct options of differential quality in some predetermined manner (e.g., writing distractors with varying degree of errors). Echternacht (1976) provided such an example wherein item writers were instructed to construct item with one correct answer, two distractors differing from the correct answer in only one aspect (one error in logic or operation) and two distractors differing from the correct answer in more than one aspect.

Still, a third scheme is to weight options a priori on the basis of a theory about the structure of the knowledge being tested. An excellent example of such application is

Smith's (1987) study on a priori weighting of vocabulary test items based on a vocabulary acquisition theory. The theory holds that learners of vocabulary make different mistakes as they progress along the continuum of word acquisition from the lowest level to the highest. Accordingly, the types of errors made by examinees can reveal their true state of knowledge and such information can be used to differentiate among examinees. Using a 50-item multiple-choice test of general English vocabulary developed specifically for this study, Smith compared the ability estimates based on Rasch dichotomous and polytomous models to determine if there were gains in validity or reliability as a result of using the polytomous scoring model rather than the dichotomous scoring model. The results indicate that the reliability and concurrent validity of the polytomous scoring ($r = 0.766$, $p = 0.823$) of a subset of items (16 items) that fit the polytomous scoring model were significantly higher than those for dichotomous scoring ($r = 0.69$, $p = 0.812$) of the same subset of items.

There is considerable amount of literature bearing on the merits of logical weighting; a summary of the earlier work is presented in Table 2.2. It can be seen from Table 2.2 that logical weighting does not exhibit consistent gains in reliability and validity, though most found internal-consistency reliability to be improved as a result of a priori weighting. With respect to validity, some studies have shown increases for logically weighted scores (Hambleton et al., 1970; Jacob & Vanderventer, 1968), whereas at least one study demonstrated a statistically significant decrease in both predictive and concurrent validity (Kansup & Hakstian, 1975).

The conflicting findings regarding a priori weighting is not totally unexpected, given that a priori weights determined by judges are far from impeccable even with the

Table 2.2 Summary of Studies Comparing the Reliability and Validity of Number Right with Logical Option Weighting Methods (N = Sample Size, n = Number of Items)

Reference	Weighting Method	Measure Compared	Results
Cross, Ross, & Geller (1980)	Judges' Ratings	Cronbach 's α Predictive Validity	Varied Across Test
Davis & Fifer (1959)	Empirical & Judges' Ratings (2 judges)	Parallel-form Reliability Predictive Validity Concurrent Validity	Increased Unchanged Unchanged
Downey (1979)	Judges' Weights (7 judges)	Hoyt Reliability Predictive Validity	Increased Increased
Echternacht (1976)	A priori Weights (By item writers)	Cronbach 's α Concurrent Validity	Decreased Decreased
Hambleton et al. (1970)	Facet Analysis Weights Judges' Weights (22 judges)	Split-half Reliability Predictive Validity	Increased Increased
Jacobs & Vanderventer (1968)	A priori Weights (Facet Analysis)	Test-retest Reliability Predictive Validity Concurrent Validity	Increased Increased Increased
Kansup & Hakstian (1975)	Judges' weights (44 judges)		
	Verbal	Cronbach's α Test-retest Reliability Predictive Validity Concurrent Validity	Increased Decreased Unchanged Unchanged
	Math	Cronbach's α Test-retest Reliability Predictive Validity Concurrent Validity	Increased Decreased Decreased Decreased
Nedelsky (1954)	Judges' Ratings	Cronbach 's α	Increased
Patnaik & Traub (1973)	Judges' Rankings (61 judges)	Split-half Reliability Predictive validity	Increased Decreased

assistance of elaborate statistical techniques checking for errors. Moreover, the latent "degree of correctness" is rather equivocal for some subject areas. As noted by Patnik and Traub (1973), it is easier to weight vocabulary item options than to do so with mathematical reasoning item options since the latent "degree of correctness" for

vocabulary item options is more readily approachable. Thus, for some content domains, the a priori weights obtained through judgmental procedures may not truly reflect the degree of correctness of the response options for which the weighting is administered.

Logical weighting lends itself readily to attitude inventories where attitude expressed in the responses can be ordered as to their difference in degree of agreement or disagreement (Gage, 1957; Yee & Kriewall, 1969). With aptitude or achievement tests, however, the items that have been used in many tests are not written with option weighting in mind and, therefore, it is not possible to determine logically which responses should be weighted most heavily. In this case, empirical weighting is brought in as a viable approach to establishing differential option weights.

2.3.3 Empirical Weighting

In contrast to logical weighting, empirical weighting assigns weights based on the responses of examinees themselves. Empirical weighting procedures involve using linear methods to obtain scoring weights for each of the response options so as to maximize the reliability and validity of the test scores. Guttman (1941) is credited with the development of an option weighting procedure that has been subsequently adapted by many researchers for use in option weighting studies. This weighting procedure computes reciprocal averages as option weights and uses these weights in an iterative process to maximize the internal consistency of the test. Hence the reciprocal averages weight has been given another name "Guttman weight" and the procedure "Guttman weighting." Guttman weighting involves the following steps:

1. Assign initial weights of 1 to correct responses and 0 to incorrect responses. Compute the total score for each examinee according to these initial weights.
2. Calculate the option mean score for each option and use it as the weight for that option. Recalculate a new option mean for each option according to the new weights.
3. The process of iteration continues until coefficient alpha stabilized according to some predetermined criterion.

Guttman's weighting method is not restricted to values determined by internal weighting of response options; external criterion can also be used to develop weights. According to Wang and Stanley (1970), Guttman also proposed using a quantitative external criterion as a basis for weighting of options; the weight for an option is the mean criterion score of persons choosing that option. Weighting in this way maximizes the correlation between the criterion scores and item scores. Guttman's weighting technique, along with its modifications (e.g., replacing option mean with mean standardized scores), has been widely used in empirical studies (Hendrickson, 1970; Raffeld, 1975).

In addition to the Guttman weighting method, other option weighting schemes have been reported in the literature. For example, Davis and Fifer (1959) weighted options in proportion to the point-biserial correlations between choosing each option and total score, thus, options chosen by examinees with higher total scores on the test receive greater weights than those chosen by examinees with lower scores. Guilford (1941) introduced a method in which the difference in proportions of upper and lower scoring groups choosing an option is used as the option weight. Bejar and Weiss (1977)

proposed a successive integer weighting system whereby successive integer weights (e.g., 4, 3, 2, 1, 0, determined by the researchers) are assigned to response options such that the best option (the correct answer) has the largest value and the worst option has the smallest value (usually set to 0). Serlin and Kaiser (1978) used the eigenvalue elements for the first principle component of the intercorrelation matrix for all test options as the weights. A summary of the option weighting methods and the resulting findings is presented in Table 2.3.

Results based on empirical weighting indicate that the use of option weighting generally increased reliability, but not validity (Davis & Fifer, 1959; Sabers & White, 1969). Hendrikson (1971) conducted a study with the Scholastic Aptitude Test using Guttman weighting method and found substantial increases in reliability and lower intercorrelations of the verbal and quantitative subtests. Reilly and Jackson (1973) reached similar conclusions using similar procedure with the Graduate Record Examination. Hendrickson (1971) suggested that option weighting purified the trait being measured on the test, which would lead to an increase in internal consistency reliability and less overlap with other measures. Consequently, option weighting has a positive effect on internal consistency reliability but seems to have mixed effect on validity. Only when the criteria and the predictor test are similar in content can empirical weighting improve predictive validity slightly (Echternacht, 1976).

It should be noted that the bulk of research on empirical option weighting was conducted in 1960s and 1970s. The inconsistent results produced by empirical studies, along with the cost and complexity of option weighting, have apparently contributed to its decline. From 1980 onwards, IRT models have been increasingly used in test

Table 2.3 Summary of Studies Comparing the Reliability and Validity of Number Right with Empirical Option Weighting Methods (N = Sample Size, n = Number of Items)

Reference	Weighting Method	Measure Compared	OP vs. NR
Bejar & Weiss (1977)	Point biserial weights	Cronbach's α	Increased
	Reciprocal averages weights		Increased
	Successive integer weights		Increased
Claudy (1978)	Biserial Correlation	Cronbach's α	Increased
	Reciprocal Averages		Increased
	Proportion Weights		Increased
Cross, Ross, & Geller (1980)	Reciprocal Averages	Cronbach 's α	Increased
		Concurrent Validity	Unchanged
Davis & Fifer (1959)	Point-biserial Correlations	Cronbach's α	Increased
		Predictive Validity	Unchanged
		Concurrent Validity	Unchanged
Downey (1979)	Reciprocal Averages	Hoyt reliability	Increased
		Predictive validity	Unchanged
Echternacht (1976)	Reciprocal Averages	Cronbach's α	Increased
		Concurrent Validity	Increased
Guttman (1954)	Reciprocal Weights	Cronbach's α	Increased
Hendrickson (1971)	Reciprocal Averages	Cronbach's α	Increased
Raffeld (1975)	Reciprocal Averages with Constant Omission Weights	Hoyt reliability	Increased
		Predictive Validity	Increased
	Reciprocal Averages with Differential Omission Weights	Hoyt reliability	Increased
		Predictive Validity	Decreased
Reilly & Jackson (1973)	Reciprocal Averages	Cronbach's α	Increased
		Parallel-form reliability	Increased
		Predictive validity	Decreased
Sabers & White (1969)	Proportion Weights	Spearman-Brown Reliability	Unchanged
		Predictive Validity	Unchanged
Serlin & Kaiser (1978)	First Principal Component	Cronbach's α	Increased

development and item analysis, which has further reduced the attractiveness of the traditional option weighting methods.

In recent years, however, there is renewed interest in option weighting fueled in part by the introduction of a new option weighting procedure (Sympson, 1988, 1990). Unlike the previous research which dealt primarily with achievement and aptitude testsand were interested in measures such as coefficient alpha, parallel form reliability, test-retest reliability, concurrent and predictive validity, recent studies focus on the effectiveness of option weighting with respect to classification accuracy related to criterion-referenced tests where pass-fail decisions are made (Sympson & Haladyna, 1988; Sympson & Davison, 1989; Haladyna, 1990). Findings from these studies suggest that empirical option weighting typically produced slightly more reliable domain score estimates and more consistent pass-fail decisions than number right scoring. These studies lend support to the contention that while there is no consistent evidence to support of the use of option weighting over number right scoring, nor are there any reasons to assume that option weighting might not improve score characteristics in specific situations.

2.4 IRT Model-based Polytomous Scoring

Item response theory is a general statistical theory that relates examinees' performance on a test to the underlying construct (e.g., ability) purportedly measured by the test (Hambleton, 1989). An item response function – often called the item-characteristic function – describes the relationship between examinee performance on each item and the ability measured by the test. Monotonically increasing, this item characteristic function provides probabilities of examinees at various ability levels answering an item correctly.

Item response theory offers a new avenue to weighting response options that deviates substantially from the more traditional linear option weighting methods. Linear option weighting methods are score-based, that is, weights are derived to maximize the correlation between an item score and the test score. This may produce undesirable side effects such as the derived weights are dependent on the calibrated items and the calibration sample. With item response theory models, however, it is possible to obtain scoring weights that are independent of the items calibrated in the set and the sample from which the weights are derived, as item response theory assumes that the characteristics of an item are independent of the ability distribution of the examinees and the characteristics of an examinee are independent of the set of test items administered (Hambleton, Swaminathan, & Rogers, 1991, p. 18).

Item response theory models that have been applied in test development, item analysis, test equating, and computerized adaptive testing are basically unidimensional models conforming to the two fundamental assumptions of IRT -- unidimensionality and local independence, which hold that an examinee's response on a set of test items is a function of the examinee's ability; and when that ability is held constant, examinee's response to any pair of items are statistically independent (Hambleton et al., 1991). These unidimensional models are further classified into two categories: dichotomous and polytomous models. Dichotomous IRT models deal with item responses that are binary, that is, they are scored as either correct or incorrect (0-1 score metric), whereas the polytomous models deal with response data scored in multiple categories. The benefit of having polytomous-response models is that by modeling the probability of selecting each response category, as opposed to only modeling the probability of

selecting correct response, potential useful information about an examinee's level of ability that is contained in the complete pattern of responses is obtained. The following sections describe the IRT models considered in this study.

2.4.1 Three-Parameter Logistic IRT Model

The three-parameter logistic IRT model is mathematically defined as follows:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

where $P_i(\theta)$ is the probability with which an examinee of ability θ answer item i correctly. b_i is the item difficulty parameter; representing the probability of an examinee with given ability θ having a 50% chance of answering item i correctly, a_i is the item discrimination parameter, c_i is the guessing parameter, representing the probability of examinees with low ability answering the item correctly, and D is the scaling factor usually set equal to 1.7 (Hambleton, Swaminathan, & Rogers, 1991).

In addition to the item characteristic function described above, another useful function is an item information function, which describe the contribution of particular items to the ability estimation at any point along the ability continuum (Hambleton, 1989). Item information function for the three-parameter logistic model is defined as follows:

$$I_i(\theta) = \frac{2.89a_i^2(1 - c_i)}{[c_i + e^{1.7a_i(\theta - b_i)}][1 + e^{1.7a_i(\theta - b_i)}]^2}$$

where θ, b_i, a_i, and c_i are defined as for the three-parameter logistic model. For the three-parameter logistic model, when $c_i > 0$, an item provides its maximum information at an ability level slightly higher than its difficulty (Hambleton et al, 1991). The test information function is simply the sum of the item information functions at θ, which takes the form:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

2.4.2 Polytomous IRT Models

A wide array of polytomous IRT models has been introduced during the last three decades. Thissen & Steinberg (1986) developed a classification framework for classifying both dichotomous and polytomous models along the measurement continuum. The classification categories identified for the polytomous models are: difference models, divide-by-total models, and the left-side added divide-by-total models. Of the three categories, models in the third category -- the left-side added divide-by-total models -- are central to this research (see Thissen & Steinberg for a detailed discussion on models included in other categories). Models included in this category are Samejima's (1979) multiple-choice model, the multiple-choice model of Thissen and Steinberg (1984), and Sympson's (1983) Model 6. This group of models is an extension to Bock's (1972) nominal model (a divide-by-total model) by adding parameters for a latent response category "Don't Know" (Thissen & Steinberg, 1984), modeling the group of examinees who are totally undecided as to which response category they should select. The nominal model and the multiple-choice models are

usually used with multiple-choice items in which it is difficult to order distractors according to their relative degree of correctness. Since there is no a priori ordering of response categories of the multiple response items used in this study, the nominal model and the multiple-choice model are considered to be the appropriate models for polytomous calibration of item responses. The properties of each model are detailed in the following sections.

2.4.3 Nominal Response Model

The nominal response model was proposed by Bock (1972). The model can be mathematically described as:

$$P_{ik}(\theta) = \frac{e^{(a_{ik}\theta + c_{ik})}}{\sum_{k=1}^{m_i} e^{(a_{ik}\theta + c_{ik})}}$$

where $P_{ik}(\theta)$ is the conditional probability that candidate with ability θ chooses choice k (k = 1, 2, 3, …, *mi*); m_i is the number of score categories. a_{ik} is the discrimination parameters for the categories, c_{ik} reflects the relative frequency with which examinees select each of the response alternatives. The probability of selecting category k is affected not only by the propensity towards k, but also by the propensities toward all other categories. The item category characteristic curves for categories with the largest *a* and the smallest *a* in an item are monotonically increasing and decreasing, respectively, as a function of θ. Categories associated with intermediate values of a

have non-monotonic response functions. Two constraints are imposed on the model to resolve the indeterminacy in the item category parameters:

$$\sum_{k=1}^{mi} a_{ik} = \sum_{k=1}^{mi} c_{ik} = 0$$

A problem regarding this model is that the notion of a monotonically decreasing response function presents a problem when applied to MC items because it assumes that as proficiency decreases the probability of selecting one particular incorrect response approaches unity and all the others go to zero (Thissen & Steinberg, 1984). Samejima proposed a solution to this problem (1979).

2.4.4 The Multiple-choice Models

Samejima extended Bock's nominal model by adding a response category labeled "zero," representing the class of examinees who randomly guesses at answers to an MC item. The model can be described as follows:

$$P_{ik}(\theta) = \frac{e^{(a_{ik}\theta + c_{ik})} + d_{ik} e^{(a_0 + c_0)}}{\sum_{k=1}^{m_i} e^{(a_{ik}\theta + c_{ik})}}$$

where k = 1, 2, … , m_j .

In Samejima's (1979) multiple-choice model, the "zero" category is a latent category consisting of both the proportion "zero" that guess at each of the observable response alternatives and those who chose those alternatives intentionally), thus, the d_k were fixed and set equal to $1/m_j$; this represents the hypothesis that those of sufficiently

low proficiency have equal probabilities of selecting at random each of the response alternatives.

Thissen and Steinberg (1984) disagreed with this assumption and modified Samejima's multiple-choice model by estimating the "zero" category, which is designated as Don't Know (DK) category in their presentation. One constraint is placed on DK to solve the indeterminacy of parameter estimation:

$$\sum d_k = 0.$$

In contrast to dichotomous models where the information function is defined at the item level, the information function for the polytomous IRT models may be estimated for each response category as well as for the item. Samejima (1969) derived the category information function for item i as:

$$I_{ix}(\theta) = \frac{[P'_{ix}(\theta)]^2}{[P_{ix}(\theta)]^2} - \frac{P''_{ix}(\theta)}{P_{ix}},$$

where $P_{ix}(\theta)$ is the probability of obtaining a category score of x for a given θ, and $P'_{ix}(\theta)$ and $P''_{ix}(\theta)$ are the first and second derivatives of $P_{ix}(\theta)$, respectively. The item information function for item i is:

$$I_i(\theta) = \sum_{x=0}^{m_i} I_{ix}(\theta) P_{ix}(\theta).$$

2.4.5 Polytomous Scoring of Multiple Category Data

The advantage of using polytomous models versus dichotomous models in the calibration of multiple category data is that polytomous models are capable of extracting information from incorrect responses and, by thus doing, they generally

increase the accuracy of ability estimation, particularly over the lower half of the ability range. This advantage is illustrated in Bock (1972) and Thissen (1976) using the nominal response model. Both studies examined the usefulness of polytomous scoring of multiple category data and found that, in terms of test information function, polytomous scoring yields from one third more to nearly twice the information of dichotomous scoring, especially for the lower half of the ability range. However, there is no substantial difference between the two scoring methods for the upper half of the ability range. This is conceivable, as pointed out by Thissen, for examinees of higher ability are less likely to select incorrect choices and accordingly there is less information available in the incorrect responses for the upper half of the ability range. Since information function is the inverse of the standard error of measurement, an increase in information simultaneously translates into a decrease in measurement error. As the standard error of measurement decreases, the accuracy with which the ability is estimated is improved. In other words, polytomous scoring improves the accuracy of ability estimation for the lower half of the ability range. Thissen further postulated that the more difficult the test is, the more incorrect responses may be expected; and the more incorrect responses that are available, the more improvement may be expected from multiple category scoring. Other studies bearing on the merits of polytomous scoring reached similar conclusions (Drasgow, Levine, Williams, McLaughlin, & Candell, 1989; Levine & Drasgow, 1983; Sympson, 1983, 1993;Thissen & Steinberg, 1984).

Huynh and Casteel (1987) evaluated the usefulness of Bock's nominal response model with respect to the validity of pass-fail decisions. They found that the use of

Bock's nominal model for moderate-length tests did not produce decisions that differ substantially from those based on raw scores and the validity of those decisions did not change noticeably when different ability estimates were used (i.e., raw scores vs. Bock ability estimates). They did observe, however, that when the test was short, the pass-fail decisions based on Bock ability estimates and those based on raw scores were in less agreement for examinees at the lower end of the ability range. Huynh and Casteel (1987) postulated that for that particular score range, the ability tapped by the Bock model differed from that implied by the raw scores.

It is clear from the foregoing discussion that polytomous scoring of multiple category data generally increase the precision of ability estimation, particularly over the lower half of the ability range, using information accrued in incorrect responses. It follows that the more information available in incorrect responses, the more improvement may be expected from polytomous scoring. The extent to which the benefits of polytomous scoring can be realized largely depends on the characteristics of the items on the test and the examinee population, it seems prudent to use polytomous scoring when the items are difficult and when very accurate ability estimates are required.

2.5 Summary

The multiple response item type has gained attention in recent years due to its flexibility in item representation and response mode that are unattainable with the more traditional type of multiple-choice item (i.e., one-answer single-response multiple-choice item). Moreover, it has been recognized that the multiple response type of items

is capable of measuring complex ability, knowledge, and skills more readily than its MC counterparts. Multiple response items also demonstrate superiority over performance assessment tasks in the accuracy and economy of scoring. The increasing use of the multiple response items in achievement, aptitude, licensure, and certification tests raises question about the appropriate scoring rules for MR items. The existing scoring techniques treat MR items either like MC items and score them as correct or incorrect, or like the multiple true-false items where each response option is scored as an independent item. Both classes of scoring methods lack the mechanism to fully assess partial knowledge, with the second class of scoring techniques having additional problem such as the dependency among response alternatives. Classical option weighting methods and polytomous IRT models have been considered to be capable of resolving these problems and thereby improving the precision of measurement.

The comprehensive review of the literature on option weighting of MC tests reveals that reliability estimates (e.g., coefficient alpha, parallel form reliability) are generally increased by the use of empirical option weighting when compared to conventional number right scoring. Although the increases in reliabilities observed in many studies have not attained statistical significance, the practical importance of the increases has much to recommend the use of option weighting in some cases. It has been recognized that option weighting can extract additional information from a given set of items. This allows the test developer to use fewer items in a test, while retaining a previously set reliability standard. This, in turn, is especially desirable in the case where items are difficult and/or expensive to construct.

In addition, studies on polytomous IRT models as they relate to multi-category data have shown that incorporating the amount of information contained in incorrect responses into ability estimation generally produces more reliable ability estimates for examinees, particular low ability examinees (Bock, 1972; Thissen & Steinberg, 1984; Thissen, Steinberg, & Mooney, 1989). This feature is especially useful in computerized adaptive testing (CAT) since many of the item selection algorithms implemented in CAT use maximum information approach, which directs the next best item administered to be the one that provides the most information at the examinee's current ability estimate, based on the responses to the previous items administered. The addition of information extracted from incorrect responses enables the item selection algorithms to reach the desired level of precision with fewer items in a shorter period of time. This hypothesis has been confirmed by De Ayala (1989, 1992), who compared CATs based on the nominal model and the three-parameter logistic model in the context of achievement testing and found that while the two models performed equally well, considerably fewer items were administered by the nominal model CAT than the three-parameter logistic model CAT. This is because the nominal response model provides more information than the three-parameter logistic model for low ability level examinees.

With respect to classification accuracy, studies have shown that empirical option weighting typically produced slightly more reliable domain score estimates and more consistent pass-fail decisions than number-right scoring, particularly in the lower half of the test score distribution (Haladyna, 1990; Sympson & Haladyna, 1988). Haladyna (1990) argued that in the context of licensure and certification testing, even small gains

in reliability are justified when such practices can improve the accuracy of classification.

In summary, option weighting and polytomous scoring using IRT models appear to improve the psychometric properties of the test scores in specific situations and it is the author's contention that these scoring methods should be applied to score MR items for optimal results when appropriate. The current study represents an attempt to determine the extent to which these scoring methods can be generalized to MR scoring. The results of the study may have some bearing on the issue of appropriate scoring rules for MR items.

CHAPTER 3

METHODOLOGY

3.1 Introduction

In this chapter, the methodology for the study is described. The purpose of the study was to compare various scoring strategies and to investigate the differences between classical and IRT scoring models as well as the differences between dichotomous and polytomous scoring models as they related to the multiple response item type. Different scoring methods were compared with respect to measurement efficiency and classification accuracy as evaluated by several criteria including reliability, item and test information functions, and accuracy of pass-fail decisions. Four scoring models, polyweighting, three-parameter logistic model, Bock's nominal model, and the multiple-choice model of Thissen and Steinberg were considered in this study. Three comparisons were made in the study:

1. Comparison between three-parameter logistic model and polytomous models (dichotomous vs. polytomous scoring).
2. Comparison between Bock's nominal model and the multiple-choice model of Thissen and Steinberg (model parsimony).
3. Comparison between polyweighting and each of the polytomous IRT models (classical vs. IRT scoring).

The item pool used, the procedures investigated, and the criteria for evaluating results are detailed in the following sections.

3.2 Item Pool

The data for this study were taken from the field test data of a large-scale computerized certification examination designed to certify entry-level technicians in computer hardware applications. The field test data consisted of six parallel linear forms of 70 items each; each form was administered to roughly 3000 examinees. Of the 70 items tested on each form, about 1/7 to 1/3 (varying from form to form) were multiple response items. These MR items measured five content domains that are essential in computer hardware applications. The number of correct answers to the MR items ranged from 2 to 4, and the exact number of correct answers was specified to examinees (e.g., "choose two"). These items were scored dichotomously on an all-or-none basis, with one point given for selecting all the correct answers and none of the distractors, and zero points otherwise.

Of the six parallel forms, two were chosen according to the following criteria: 1) the selected test consisted of a large number of multiple response items; and 2) preliminary analysis indicated that the data were unidimensional. Table 3.1 presents the general item statistics for the two samples. It should be noted that these statistics were computed by scoring multiple response items dichotomously on an all-or-none basis.

Table 3.1 General Characteristics of the Tests

Form	Number of Examinees	Number of Items	Mean Total Test Scores	SD	Percentage Passing	Passing Score (theta)	Coefficient alpha
A	2754	70	47.66	10.09	60.1%	-0.19	.89
F	2718	70	48.37	9.85	62.7%	-0.27	.88

Note. Two forms were used in this study to evaluate the stability of the results.

Since one of the goals of the study was to examine how item format factors may affect examinee performance, comparisons between the MC and MR items were made to illustrate the effect. Table 3.2 presents the classical item statistics for each form. As with Table 3.1, these statistics were computed by scoring multiple response items dichotomously on an all-or-none basis, that is, full points were given if all the response options in the MR items were answered correctly (i.e., correct answers and none of the incorrect answers were selected for an item), and zero points otherwise (including blanks and partially correct answers).

It can be seen from Table 3.2 that MR items were, on average, more difficult, more discriminating, and required more time to finish (45% more for Form 1; 13% more for Form 2) compared to MC items. This is consistent with Hsu et al.'s (1984) findings that MR items are more difficult compared to their MC counterparts.

With respect to the average time used, it is clear that considerably longer responding times were required for MR items than for MC items. From the perspective of cognitive psychology (Britton, Glynn, Meyer, & Penland, 1982), this result indicated that more of the examinee's cognitive capacities were required for MR items than for MC items. This increase in cognitive processing demand could imply more reading by the examinees or could be indicative of higher-level processing while working on the item. This added an empirical facet to Dressel and Schmid's (1953) hypothesis that MR item may provoke more extended thought process than other item formats, as reflected by the longer time allocated to responding to the task.

Table 3.2 Comparison of Multiple-choice and Multiple-response Items (N = number of items)

Form	Item Type	N	Item Difficulty (p)			Item Discrimination (r)			Time Used (In seconds)		
			Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
1	MC	50	.73	.19	.13 - .96	.34	.11	.10 - .56	38	14	19 - 84
	MR	20	.61	.19	.27 - .94	.39	.09	.22 - .48	53	23	33- 123
	Overall	70	.69	.19	.13 - .96	.36	.10	.10 - .56	42	18	19- 123
2	MC	52	.70	.17	.15-.96	.37	.13	.03-.60	38	14	15- 69
	MR	18	.65	.21	.33-.98	.42	.11	.09-.59	48	12	27- 74
	Overall	70	.69	.18	.15-.98	.38	.13	.03-.60	41	14	15- 74

3.3 Scoring Methods under Consideration

Both dichotomous and polytomous scoring systems were investigated in this study. The polytomous scoring systems were further divided into two categories: classical and IRT scoring. The following sections present the characteristics of each of the scoring systems.

3.3.1 Dichotomous Scoring

The scoring system for the multiple response items (as well as the multiple-choice) in the test was to give one point for a correct answer (i.e., all of the correct answers and none of the distractors are selected) and zero points for incorrect answers (including omits and partially correct answers). Thus, the operational number right scoring system became the baseline system for making comparisons.

It should be noted that scoring each response option of the MR items as right or wrong (i.e., treating each option as a true-false item) was not an option for the two datasets because of the dependence across response options (e.g., if examinees have to choose two, obviously the others cannot be chosen).

3.3.2 Polyweighting

A linear option weighting procedure, polyweighting (Sympson, 1993), was used to score multiple response items polytomously. Polyweighting is chosen for its known robustness to some of the problems associated with other linear weighting methods. As discussed in Chapter 2, empirical weighting methods assign differential weights to response options on the basis of the option's attractiveness, average standardized score

of examinees selecting an option, the correlations between choosing each option and total score, as well as other similar procedures. These weighting procedures typically seek to maximize internal consistency reliability of the test.

As with any linear regression techniques, these linear methods have two major disadvantages: first, option weights derived from these procedures are sample dependent; that is, weights derived from one sample of responses would differ from that obtained from another sample of responses. Hence, it is critical that weights derived in a particular sample be cross-validated to avoid capitalizing on the idiosyncrasy of the sample from which the weights are obtained. The second problem is that these weights are linearly dependent of the difficulty of other items on the test. If an item is calibrated along with a set of easy items, the obtained scoring weights will be different than if the item were calibrated along with a set of difficult items (Sympson & Haladyna, 1988). A corollary to this rule is that weights assigned to incorrect answers often exceed that assigned to the correct answers.

Polyweighting overcomes the second problem by assigning weights based on the percentile ranks of examinees choosing an option; the scoring weight assigned to each response option is approximately equal to the mean percentile rank of examinees choosing that option in the item calibration sample. The procedure assigns scoring weights as follows:

(1) Compute each examinee's proportion correct score among items that were administered to the examinee.

(2) Convert the examinee's proportion correct score to percentile rank relative to those examinees who were administered the same item set.

(3) For each item, determine the mean percentile rank among examinees who chose each possible response category. Round the mean percentile rank to the nearest integer and use it as initial polyweights.

(4) Compute provisional polyscore for each examinee. This polyscore is equal to the mean of the polyweights of the categories chosen by the examinee. Convert the polyscore to percentile rank relative to those examinees who were administered the same item set.

(5) Continue the iteration by using the polyscores from Step 3 as initial weights to recalculate new polyscore for each examinee. The process of iteration stops when the mean squared correlation ration between items and percentile ranks stops increasing.

According to Sympson and Haladyna (1988), the use of mean percentile rank in polyweighting is equivalent to equipercentile equating of proportion correct scores from different item sets, thereby solving the problem of variation in item difficulties. As a result, polyweighting produces scoring weights for a given item that are independent of the difficulty of other items in the analysis. Moreover, the scoring weights are bounded so that an examinee can never receive more credit for an incorrect response than for a correct response.

In addition to the advantages mentioned above, polyweighting works well with small samples and is unaffected by the dimensionality of the calibration data. Because of these advantages, polyweighting has been used in several studies to compute option weights for response categories (Blankenship, Cesare, & Sympson, 1992; Davey,

Godwin, & Mittelholtz, 1997; Sympson & Haladyna, 1988; Sympson & Davison, 1989). For the same reasons, polyweighting was used in this study.

3.3.3 IRT Model-based Polytomous Scoring

Nonlinear polytomous scoring of multiple response items was performed by polytomous IRT models. Polytomous IRT models provide a way to investigate individual differences related to response category selection. In contrast to dichotomous IRT models that model the probability of selecting the correct answer; polytomous IRT models model the probability of selecting each response category. Since there is also likely to be information conveyed in the incorrect responses, modeling all of the category responses can improve the accuracy of ability estimation. Moreover, polytomous IRT modeling provides rich diagnostic information about examinee cognition that is not apparent from total test scores (Mislevy, 1995).

Polytomous IRT scoring also can produce sample-free item parameter estimates and item-free person parameter estimates, given that the set of items calibrated with a polytomous IRT model is unidimensional, and that the chosen model fits the data. Polytomous IRT models considered in this study were Bock's nominal response model, and the multiple-choice model of Thissen and Steinberg. Both models have been used for polytomous calibration of multiple-choice items where the order of the response category is nominal, that is, the category cannot be ordered to represent varying degree of the trait measured by the item (Bock, 1972; Thissen & Steinberg, 1984).

3.3.4 Procedures

Two test forms (Forms 1 and 2) were selected for this study. The general statistics for each form are detailed in Table 3.1. Including two forms in the analysis provided a viable means to examine whether the results replicate over different forms. If similar results were obtained from different analyses, then we may be able to generalize the findings to other tests that contain MR items.

The data for this study consisted of examinees' responses to 70 items. The item responses were dichotomously scored and the number right scores were used as the baseline for comparisons. Since there is likely to be "shrinkage" in the amount of reliability observed when polyweights are applied in new samples (Blankenship et al., 1992), it was deemed essential to cross-validate the polyweights obtained in a calibration sample against one sample to which the polyweights are applied. In order to do so, the data were randomly split into two halves, one constituting the calibration sample and the other application sample. Polyweights were computed from the calibration sample and were applied to the application sample. All comparisons were made based on calibrations of the application sample.

Since both the nominal model and multiple-choice model were considered in this study, it was deemed practical to compare the performance of two polytomous IRT models first. For this purpose, the test was calibrated twice, once with the nominal model, and once with multiple-choice model, using the raw response patterns; the item and test information functions obtained from the two calibrations were compared and contrasted.

For the comparison between dichotomous and polytomous scoring methods, the multiple response items were scored dichotomously and calibrated along with the rest of the multiple-choice items using the three-parameter logistic model. The resulting item and test information functions were compared to the information functions obtained under the polytomous IRT models.

For comparison between classical and IRT polytomous scoring, the test was scored by polyweighting procedure. Scoring weights derived from the calibration sample were used to score item responses in the application sample and the resulting scores and the option weights were used for comparisons between linear and IRT polytomous scoring methods.

The computer program POLY (Sympson, 1990) was used for the polyweighting analysis. MULTILOG was used for both the dichotomous and polytomous calibrations of the multiple response items since it is the only widely available program that implements parameter estimation algorithms for both the nominal model and the multiple-choice model. MULTILOG implements a marginal-maximum-likelihood algorithm to estimate parameters (Thissen, 2002).

One concern about the comparison of item statistics obtained from different calibrations is whether they are compatible with each other, that is, whether they are on the same scale. Research has indicated that when the same data are calibrated by different models, the results are comparable, the observed differences in parameter estimates, if any, are most likely be present in the guessing parameter, the slope and threshold parameters are not affected by different models (Thompson & Pommerich, 1996). In light of the findings, equating was deemed unnecessary in this analysis.

3.4 Evaluation Criteria

The effectiveness of the scoring method was evaluated by several indices, including the internal consistency and marginal reliabilities, item and test information functions, and the proportion of examinees "correctly" classified by each procedure.

3.4.1 Measurement Efficiency

Measurement efficiency was evaluated using an index of "relative information." This index is based on the Spearman-Brown formula (Brown & Thomson, 1925), which predicts the reliability of a lengthened test as a function of the initial reliability of the test. The formula can be rearranged to determine how much a given test would have to be increased in length in order to obtain a specified level of reliability (Nishisato, 1980, p. 118). The formula is given as follows:

$$h = \frac{a_s(1-a)}{a(1-a_s)}$$

where a is the reliability of the unweighted scores (i.e., the conventional number right scores), and a_s is the reliability of the weighted (i.e., option-weighted) scores. The statistic, h, indicates how much the unweighted test would have to be increased in length in order to obtain the reliability observed in its weighted counterpart.

Another measure of relative efficiency is the ratio of test information functions obtained under different scoring methods (Hambleton et al., 1991, p. 94). Test information function is the IRT equivalent to reliability as is defined in classical test theory and therefore, the ratio of test information functions may be used as an indicator of the relative measurement efficiency afforded by different IRT models. In fact, the ratio of information functions is a better measure of relative efficiency because it allows

for further examination of the gains in information at different ability levels that might not be possible with the classical reliability statistic (although procedures have been developed to compute conditional standard error of measurement, it is relatively difficult to compare the statistics because of the sample- and test-dependency problems in classical test theory models). Hence, while the change in overall reliability from using polytomous scoring method is not significant, there might be substantial increase in reliability for the lower part of the ability range accrued from using information in incorrect responses, as research has shown that modeling incorrect responses generally resulted in moderate gains in information for low to slightly above average abilities (Bock, 1972; Thissen, 1976; Thissen & Steinberg, 1984). On this basis, it was expected that test information functions for low abilities obtained under polytomous IRT modeling would be greater than that obtained with the dichotomous IRT model.

3.4.2 Classification Accuracy

Since there was no external criterion against which the accuracy of pass-fail classifications made under each scoring scheme could be validated, an alternative criterion was sought to evaluate the accuracy of classifications made under different scoring schemes. This criterion required recalibrating only the MC items on the test and using the passing rate obtained as a baseline for comparison. To determine the passing rate, the theta value (i.e., the ability estimate in IRT) that corresponded to the operational passing score (i.e., the cutoff score obtained under dichotomous scoring) was obtained and this theta value was then treated as the true passing value and used to classify examinees as passing or failing. Since the MC items measured the same content

as the MR items and since an examinee's ability estimate is independent of the set of items that are administered to him/her, the use of a uniform theta value to classify examinees under different scoring schemes is warranted (Keller, Swaminathan, & Sireci, in press). To gauge the level of classification agreement among the different MR scoring methods, kappa (percent agreement corrected for chance, Cohen, 1965) was also computed.

3.4.3 Congruence of Weighting

Option weights obtained under different scoring models were contrasted. For each response category, the polyweight is equal to the mean percentile rank among examinees choosing the category, rounded to the nearest integer while the IRT weight is a_k, the slope (discrimination) parameter for that response category. Linear correlation between the two sets of weights was calculated and used as an index of congruence between different weighting systems. Since both sets of weights were considered optimal it was expected that they be highly correlated.

CHAPTER 4

RESULTS AND DISCUSSION

The results of this study are organized according to the evaluation criteria outlined in Chapter 3. Measurement efficiency, classification accuracy, and congruence of weighting obtained under each of the scoring models[1] outlined in the previous chapter are presented.

Since the data were split into two halves in the polyweighting analysis to cross-validate the weights obtained from one sample against those obtained from another sample, sample test statistics were also computed in the item response theory analyses. Both population and sample level results are provided. Population results were compiled over the population that took the two tests while sample level results were compiled for each of the split-half samples. These results are detailed in the following sections.

4.1 Measurement Efficiency

Measurement efficiency is evaluated using an index of "relative information," which is defined as the ratio of the reliability of a option-weighted test as a function of the initial reliability of the test in the classical test theory model, and as the ratio of the test information of the polytomous scored tests as a function of the initial information function of the dichotomously scored tests.

---

[1] MTF scoring was excluded from the study because preliminary analysis indicated a poor fit to the data.

4.1.1 Comparison of Reliability

Both coefficient alphas and marginal reliabilities are reported in Tables 4.1 and 4.2. It should be noted that marginal reliability was not available for polyscores because it is the IRT-equivalency to the internal consistency reliability for the classical test theory test scores. As a non-IRT weighting model, polyweighting does not produce marginal reliability estimates for polyscores. As discussed in the previous chapter, measurement efficiency is evaluated using an index of "relative information," the h-statistic, which predicts the reliability of a lengthened test as a function of the initial reliability of the test. A positive *h* statistic indicates the extent to which the dichotomously scored test would have to be increased in length to achieve the reliability observed in its polytomously scored counterpart. A value less than one would suggest an advantage for dichotomous scoring.

Table 4.1 Coefficient Alpha and Marginal Proficiency across Weighting Methods (Form A)

		Marginal reliability	h-statistic
3 PL Model	Population	.90	1.00
	Sample1	.90	1.00
	Sample2	.90	1.00
Nominal Model	Population	.90	1.00
	Sample1	.91	1.01
	Sample2	.91	1.01
Multiple-choice Model	Population	.91	1.01
	Sample1	.91	1.01
	Sample2	.91	1.01
Poly-weight	Sample1	.91*	1.01
	Sample2	.91*	1.01

Note. * Coefficient alpha.

Table 4.2 Coefficient Alpha and Marginal Reliability across Weighting Methods (Form F)

		Marginal Reliability	h-statistic
3 PL Model	Population	.90	1.00
	Sample1	.90	1.00
	Sample2	.90	1.00
Nominal Model	Population	.90	1.00
	Sample1	.90	1.00
	Sample2	.90	1.00
Multiple-choice Model	Population	.91	1.01
	Sample1	.91	1.01
	Sample2	.91	1.01
Poly-weight	Sample1	.91*	1.01
	Sample2	.91*	1.01

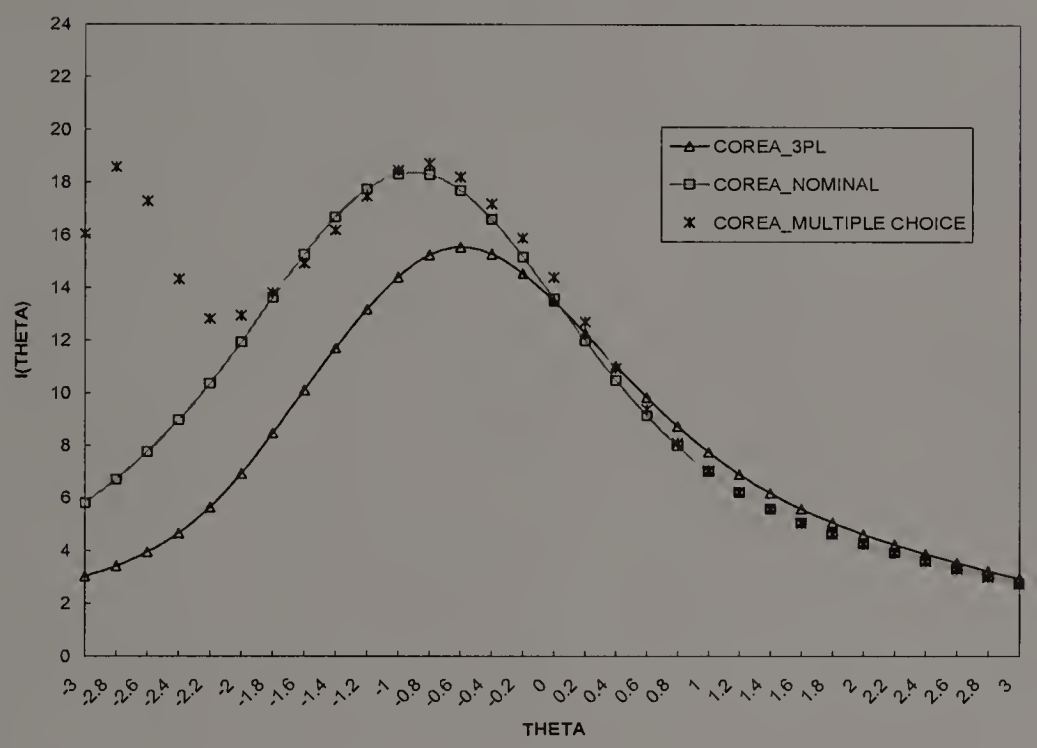
Note. *Coefficient alpha.

The h statistics from Tables 4.1 and 4.2 are all equal to or greater than one, indicating that polytomous scoring produced equally or slightly higher reliabilities than comparable dichotomous scoring of the same forms. However, the differences in reliability are trivial at best, revealing that polytomous scoring did not significantly improve the efficiency of measurement in this case, probably because 2/3 of the items were MC items and scored the same way.

4.1.2 Comparison of Test Information

Graphical illustrations of the overall test information functions for the two tests are presented in Figures 4.1, 4.2 and 4.3.

a)



b)

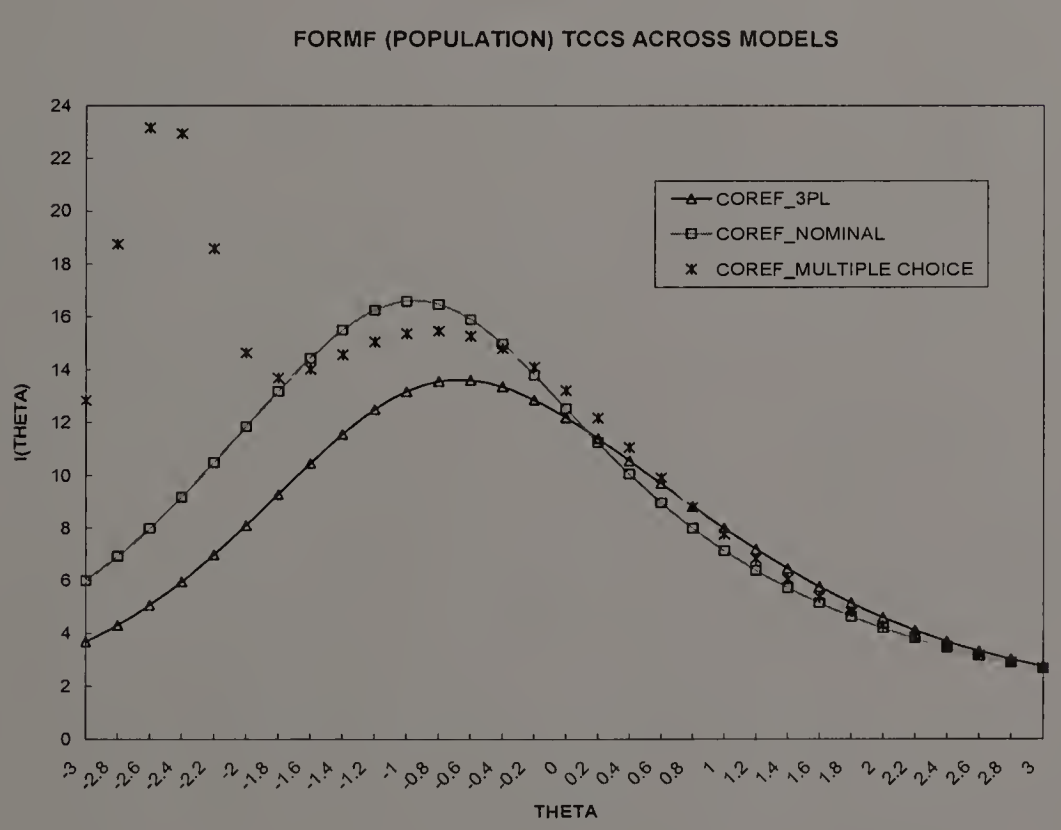
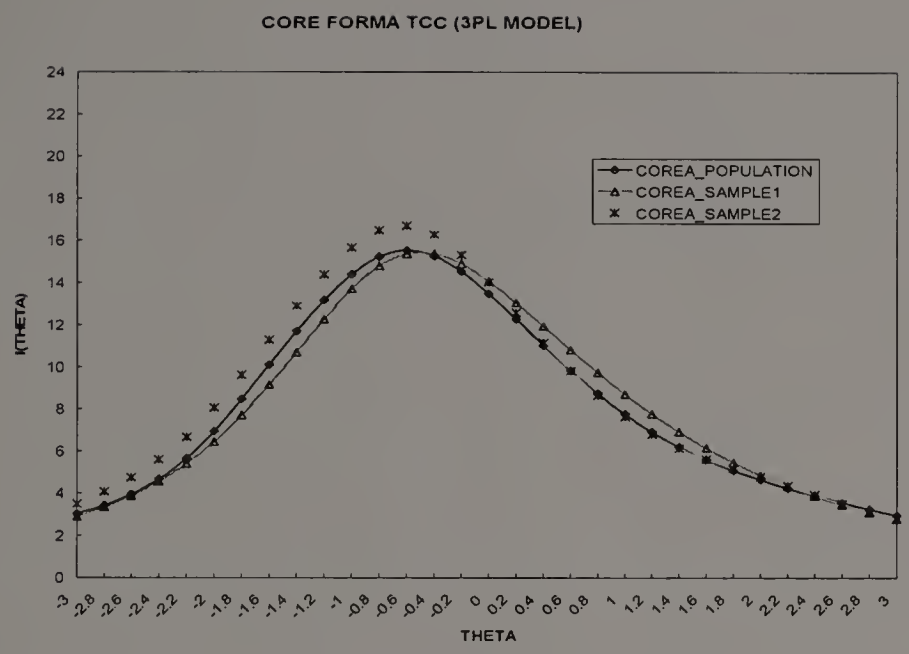
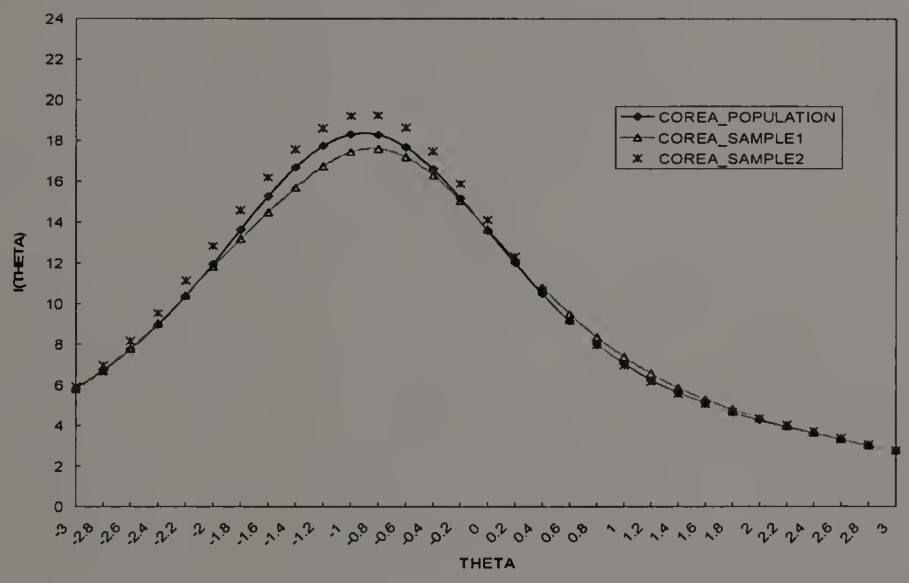


Figure 4.1 Test Information Functions across Models (Population only)

a)



b)



c)

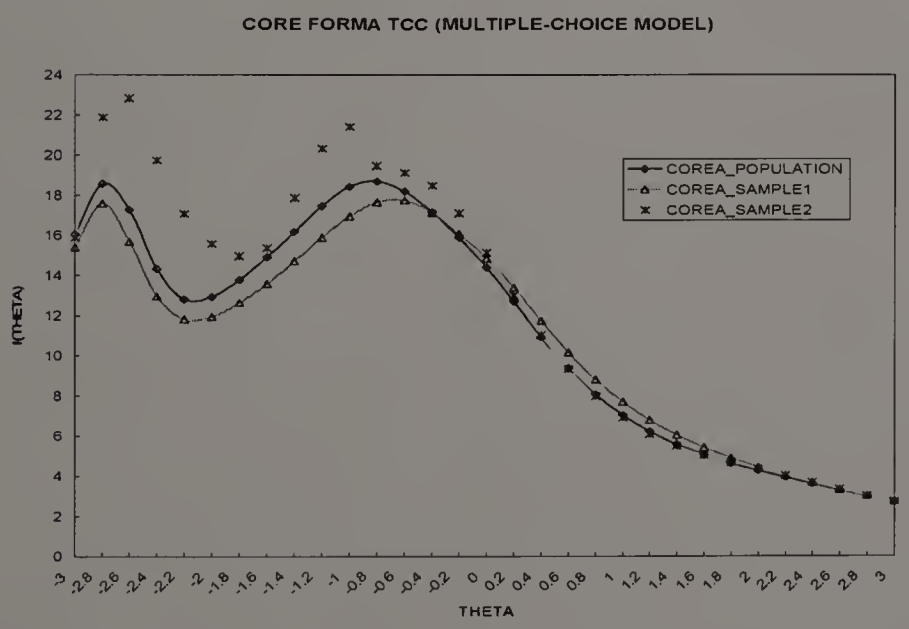
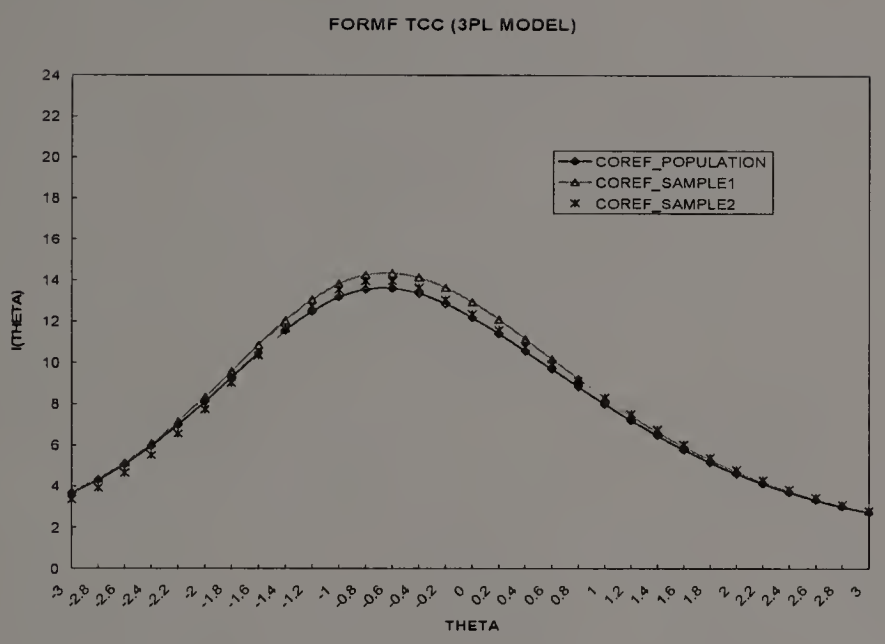
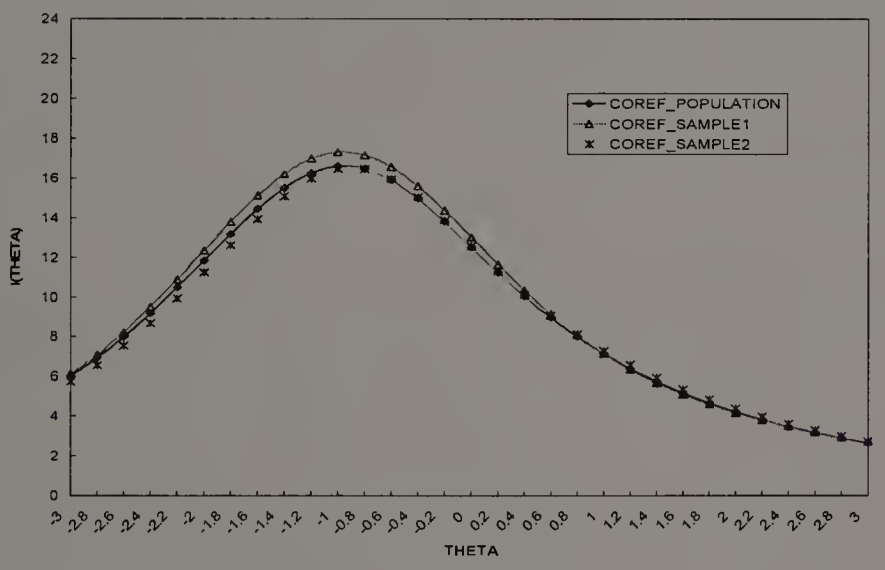


Figure 4.2 Test Information Functions within Models (Form A)

a)



b)



c)

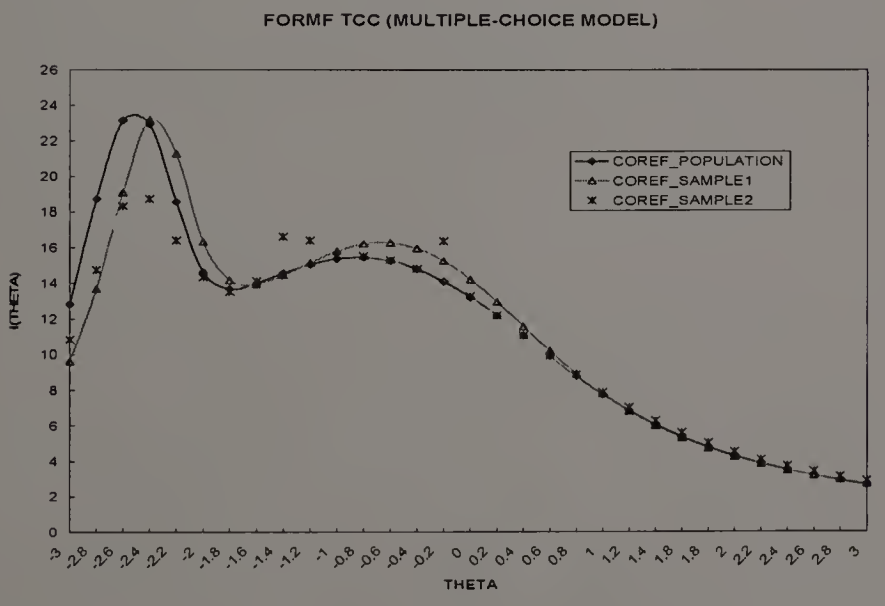


Figure 4.3 Test Information Functions within Models (Form F)

Figure 4.1 provides the test information functions for the population obtained under different scoring models. Figure 4.1 shows that when MR items were scored polytomously, test information functions increased noticeably; also, the points on the proficiency scale at which the information functions attained their maximum shifted to the left, indicating that the tests now provided the most information at the proficiency levels lower than the levels obtained under dichotomous scoring. This is consistent with Samejima (1976) and Donohue (1994)'s findings that polytomous scoring yields considerably more IRT information than does the optimal dichotomization of the same items.

Figures 4.2 and 4.3 illustrate the test information functions obtained for both the population and sample calibrations within each model for the two tests. It is obvious that the three calibrations produced test information functions that are almost identical to each other, a clear indication of the equivalency of the split-half samples and of the samples with the population. It should be noted, however, that in multiple-choice model calibration, the test information function of the Sample2 group, albeit not significant, deviated from the population and Sample1 group, indicating that the Sample2 group is less homogeneous with the other two groups.

As discussed in the previous section, local maxima were apparent in the multiple-choice model calibrations. This finding suggests that modeling incorrect responses generally results in moderate, and in this case, substantial gains in information for the lower part of the proficiency range (Bock, 1972; Thissen, 1976; Thissen & Steinberg, 1984).

Tables 4.3 and 4.4 present the test information functions for the two tests.

Table 4.3 Test Information Functions across Models (Form A)

		θ_{max}	I_θ	I_θ at the Passing Score (theta = -0.27)
3 PL Model	Population	-0.6	15.543	14.543
	Sample1	-0.4	15.390	14.909
	Sample2	-0.6	16.726	15.343
Nominal Model	Population	-1.0	18.310	15.178
	Sample1	-0.8	17.611	15.062
	Sample2	-0.8	19.238	15.892
Multiple-choice Model	Population	-0.8*	18.705*	15.894
		-2.6*	18.581*	
	Sample1	-0.6*	17.760*	16.103
		-2.8*	17.595*	
	Sample2	-1.0*	21.428*	17.111
		-2.6*	22.844*	

Note. *Bimodality. Test information functions peak at two different points on the proficiency scale.

Table 4.4 Test Information Functions across Models (Form F)

		θ_{max}	I_θ	I_θ at the Passing Score (theta = -0.19)
3 PL Model	Population	-0.6	13.617	12.866
	Sample1	-0.6	14.367	13.654
	Sample2	-0.6	13.973	13.078
Nominal Model	Population	-1.0	16.604	13.812
	Sample1	-1.0	17.310	14.369
	Sample2	-1.0	16.464	13.862
Multiple-choice Model	Population	-0.8*	15.471*	14.109
		-2.6*	23.158*	
	Sample1	-0.6*	16.293*	15.268
		-2.4*	23.189*	
	Sample2	-1.4*	16.630*	16.383
		-2.4*	18.748*	

Note. *Bimodality. Test information functions peak at two different points on the proficiency scale.

Tables 4.3 and 4.4 summarize the information functions for both the population and the two samples obtained with different scoring models. When the MR items were scored dichotomously, the information functions for each of the tests attained their maximum at the proficiency level of -0.6 for both tests, indicating that the two tests provided the most information about examinees of moderately low proficiency. When the MR items were scored polytomously, the information functions for the two tests

registered a shift to the left (-0.8 to -1.0) on the proficiency scale, indicating that the tests provided the most information about examinees at an proficiency that is lower than the proficiency obtained under dichotomous scoring. Since an item provides its maximum information at an proficiency level slightly higher that its difficulty, this finding supports the previous finding that, polytomous scoring of MC (and MR items, for this matter) items often results in an increase in test information function, as well as a decrease in item difficulty (Donoghue, 1994).

Of special interest is the information function obtained in multiple-choice model calibrations. Apparently, the model resulted in bimodality, with information functions peaked at two different points on the proficiency scale. A possible explanation for this local maxima phenomenon is that with the multiple-choice model, guessing of low proficiency examinees is modeled such that there is likely to be more information accrued from incorrect responses at the lower end of proficiency scale.

4.2 Classification Accuracy

In the credentialing testing situation, the test scores are used to make mastery classification decisions. The accuracy of classification decisions based on test scores is evaluated against a criterion, in most cases, an external criterion. Since no such external criterion existed in this study, an alternative criterion was sought to evaluate the accuracy of classifications made under different scoring schemes. This criterion requires recalibrating only the MC items on the test and using the passing rate obtained as the criterion for classifying examinees as passing or failing. The accuracy of

classification decisions based on different scoring schemes was evaluated against this criterion classification.

In order to examine the accuracy of classification, examinees were classified as "pass" or "fail" in three different ways: (1) based on MC items only (criterion classification), (2) based on MC and dichotomously scored MR items (dichotomous scoring), and (3) based on MC and polytomously scored MR items (polytomous scoring).

Examinees were classified based on the estimates of their proficiency, theta, obtained from each calibration. To classify the examinees as passing or failing, the expected theta that corresponds to the operational passing score was used as the criterion. It should be noted that the criterion passing score-- theta was based on only the multiple-choice items. Each method of scoring the MR items was then compared to the criterion classification, and a judgment was made regarding which scoring method was most consistent with the criterion classification.

Classification consistency is measured by Kappa, which is the percent agreement corrected for chance (Cohen, 1960). Kappa was computed as follows:

$$K(\lambda) = \frac{\lambda_a - \lambda_c}{\lambda_a - \lambda_c}$$

where λ_a is the actual value of the measure, λ_p the value under perfect agreement, and λ_c the chance value.

Table 4.5 reports the Kappa associated with the classification decisions.

Table 4.5 Classification Accuracy using Dichotomous Scoring of MC Items (Form A)

a) 3 PL Model vs. Criterion

		Criterion Classification	
		Fail	Pass
Dichotomous Classification (3PL Model)	Fail	922 (0.34)	156 (0.06)
	Pass	80 (0.03)	1557 (0.57)

Note. Kappa = 0.816. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

b) Nominal Response model vs. Criterion

		Criterion Classification	
		Fail	Pass
Polytomous Classification (Nominal Response Model)	Fail	867 (0.32)	98 (0.04)
	Pass	135 (0.05)	1615 (0.59)

Note. Kappa = 0.814. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

c) Multiple-choice model vs. Criterion

		Criterion Classification	
		Fail	Pass
Polytomous Classification (Multiple-choice Model)	Fail	856 (0.32)	75 (0.03)
	Pass	146 (0.05)	1638 (0.60)

Note. Kappa = 0.822. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

Table 4.6 Classification Accuracy using Dichotomous Scoring of MC Items (Form F)

a) 3PL Model vs. Criterion

		Criterion Classification	
		Fail	Pass
Dichotomous Classification (3PL Model)	Fail	963 (0.35)	133 (0.05)
	Pass	137 (0.05)	1521 (0.55)

Note. Kappa = 0.796. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

b) Nominal Response Model vs. Criterion

		Criterion Classification	
		Fail	Pass
Polytomous Classification (Nominal Response Model)	Fail	750 (0.27)	33 (0.01)
	Pass	350 (0.13)	1621 (0.59)

Note. Kappa = 0.695. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

c) Multiple-choice Model vs. Criterion

		Criterion Classification	
		Fail	Pass
Polytomous Classification (Multiple-choice Model)	Fail	962 (0.35)	205 (0.07)
	Pass	138 (0.05)	1449 (0.53)

Note. Kappa = 0.743. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

Tables 4.5 and 4.6 report the false positive and false negative errors associated with each scoring method as evaluated against the criterion classification. The discrepancy rates -- between the model and criterion classification, indicated by false positive and false negative errors -- range from 0.03 to 0.13 across various comparisons, the largest was found with the nominal model scoring in Form F (rate = 0.14). When evaluated against the criterion classification, the results are inconsistent across forms. For Form F, dichotomous scoring (0.10) showed slightly lower discrepancy rate compared to polytomous scoring (0.14 for the nominal response model, and 0.12 for the multiple-choice model). Whereas for Form A, dichotomous scoring (0.09) produced identical or slightly higher discrepancy rate compared to polytomous scoring (0.09 for the nominal response model, and 0.08 for the multiple-choice model). These inconsistent results warrant further study. It should be noted that in this comparison, only data from the population were used.

Classification similarities across models are reported in Tables 4.7 and 4.8. Inconsistency is observed in this set of comparisons. For example, the discrepancy rate – the rate between the model and criterion classification, as indicated by the false positive and false negative errors -- for the nominal response model versus multiple-choice model is 0.026 in Form A, but increases dramatically for the same comparison in Form F, to a value of 0.14. Upon further inspection, it is apparent that the discrepancy rates tend to be larger with nominal model scoring in Form F.

Table 4.7 Classification Similarities across Models (Form A)

a) Nominal Response Model vs. 3 PL Model

		Dichotomous Classification (3PL Model)	
		Fail	Pass
Polytomous Classification (Nominal Response Model)	Fail	903 (0.33)	62 (0.02)
	Pass	175 (0.07)	1575 (0.58)

Note. Kappa = 0.814. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

b) Multiple-choice Model vs. 3 PL Model

		Dichotomous Classification (3PL Model)	
		Fail	Pass
Polytomous Classification (Multiple-choice Model)	Fail	895 (0.33)	36 (0.01)
	Pass	183 (0.07)	1601 (0.59)

Note. Kappa = 0.828. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

c) Multiple-choice Model vs. Nominal Response Model

		Polytomous Classification (Nominal Response Model)	
		Fail	Pass
Polytomous Classification (Multiple-choice Model)	Fail	913 (0.34)	18 (0.006)
	Pass	52 (0.02)	1732 (0.64)

Note. Kappa = 0.943. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

Table 4.8 Classification Similarities across Models (Form F)

a) Nominal Response Model vs. 3PL Model

		Dichotomous Classification (3PL Model)	
		Fail	Pass
Polytomous Classification (Nominal Response Model)	Fail	772 (0.28)	11 (0.004)
	Pass	324 (0.12)	1647 (0.60)

Note. Kappa = 0.733. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

b) Multiple-choice Model vs. 3PL Model

		Dichotomous Classification (3PL Model)	
		Fail	Pass
Polytomous Classification (Multiple-choice Model)	Fail	1016 (0.37)	151 (0.05)
	Pass	80 (0.03)	1507 (0.55)

Note. Kappa = 0.827. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

c) Multiple-choice Model vs. Nominal Response Model

		Polytomous Classification (Nominal Response Model)	
		Fail	Pass
Polytomous Classification (Multiple-choice Model)	Fail	783 (0.28)	384 (0.14)
	Pass	0 (0.00)	1587 (0.58)

Note. Kappa = 0.702. The top number is each cell represents the number of candidates. Numbers in parentheses are the respective proportions.

The percent agreement corrected for chance index, Kappa, is higher within the polytomous scoring (nominal vs. multiple-choice) model than between the polytomous and dichotomos scoring (3pl vs. nominal, and 3pl vs. multiple-choice) model in Form A, but the pattern reverses in Form F, with Kappa lower within the polytomous scoring (nominal vs. multiple-choice) model than between the polytomous and dichotomous scoring model (3pl vs. nominal, and 3pl vs. multiple-choice). This inconclusive result warrants further study.

4.3 Congruence of Weighting

Congruence of weighting was evaluated by the correlations between different sets of weights obtained under different scoring methods.

4.3.1 Option Weights

For MR items with a number of response options (alternatives), the point associated with each response option may be determined subjectively by expert judgment or empirically based on item analysis. Option weighting based on item analysis is called empirical option weighting. A number of empirical option weighting methods exist in the measurement literature. Studies have found that empirical option weighting methods generally improve the internal-consistency proficiency of the tests (Bejar & Weiss, 1977; Claudy, 1978).

Using nominal response and the multiple-choice models, empirical weights for each of the response options of the MR items were derived. Appendices A, B, C, and D

summarize the MULTILOG estimates of item category response functions (ICRF) for the MR items in the two tests.

As defined in the previous chapter, a_k is the discrimination parameters for the categories, it could be viewed as the empirical weights assigned to the options; c_k reflects the relative frequency of the selection of each response option. Multiple-choice model appends an additional DK category, which models the probabilities of those examinees who do not know the answer who select at random each of the response alternatives.

As would be expected, for each MR item, the option with the largest value of *a* has a monotonically increasing response function; without exceptions, it is the correct answer. The option with the lowest value of a has a monotonically decreasing response function; generally, it is the least attractive option. Options with intermediate values of a have nonmonotonical functions, it appears that as examinees' proficiency increase, their probabilities of selecting these options decrease.

Analysis of the option weights using nominal response and multiple-choice models provides useful information about the properties of each of the response options of an MR item and may lead to the optimal assignment of weights to each option. This would allow a more stringent scoring rubric to be established to improve the quality of the measurement. The analysis also has potential application in test construction; as the rich diagnostic information it provides about each item could help item writers improve the quality of the items on a test.

4.3.2 Congruence of Option Weights

Tables 4.9 and 4.10 report the correlations of option weights obtained in different calibrations. Graphical presentation of the correlations can be found in Figures 4.4 and 4.5.

It can be seen from Tables 4.9 and 4.10, as well as from Figures 4.4 and 4.5, that the correlations of option weights obtained for the population and samples within each model are, not surprisingly, high, an indication of the equivalency of the split-half samples, and of the population and samples. An exception to this, however, is observed in the multiple-choice model scoring case, in which the correlations of option weights for both tests are not as high as would be expected. A possible explanation is that the parameters (i.e., a_k) estimated by the multiple-choice model are less stable, thus variations among the population and samples were observed. This is our reason to prefer the nominal model over the multiple-choice model in this study.

Also, the correlations of option weights between the polyweighting model and the nominal response model are fairly high, indicating that comparable results can be obtained with both polytomous scoring methods. The correlations between nominal response and multiple-choice model, and the correlations between polyweighting and multiple-choice model, are significantly lower compared to the correlations between polyweighting and nominal response model. As discussed above, parameter estimates produced by multiple-choice model are less stable and therefore less comparable to those obtained through two other polytomous scoring models.

Table 4.9 Correlations of Option Weights across Models (Form A)

Correlations		Nominal Response Model			Multiple-choice Model			Polyweighting	
		Population	Sample 1	Sample2	Population	Sample1	Sample 2	Sample 1	Sample 2
Nominal Model	Population	1							
	Sample1	.989	1						
	Sample2	.993	.965	1					
Multiple-choice Model	Population	.760	.772	.740	1				
	Sample1	.751	.767	.726	.869	1			
	Sample2	.692	.698	.677	.929	.787*	1		
Poly-weight	Sample1	.898	.907	.876	.674	.683	.582	1	
	Sample2	.918	.900	.919	.684	.685	.606	.972	1

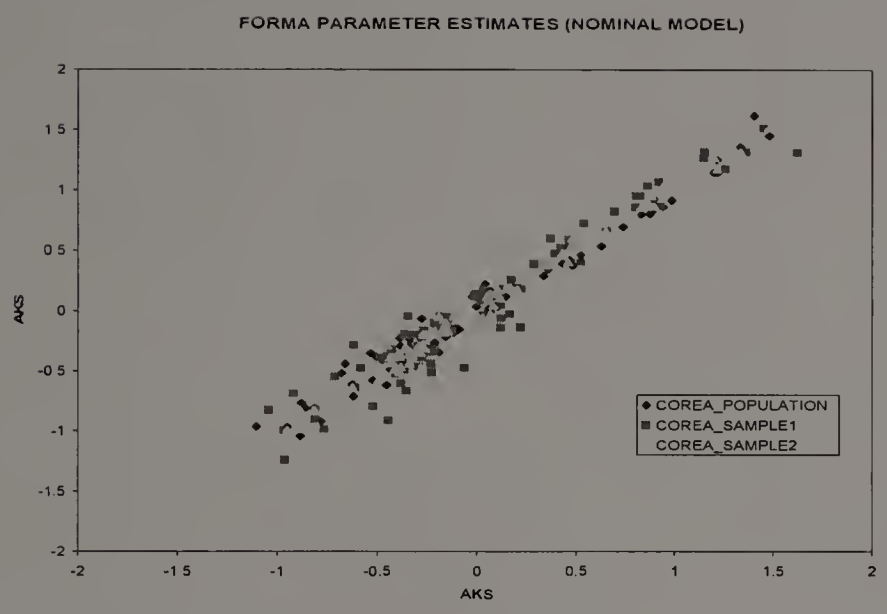
Note. *Low correlation for the two split-half samples within a model.

Table 4.10 Correlations of Option Weights across Models (Form F)

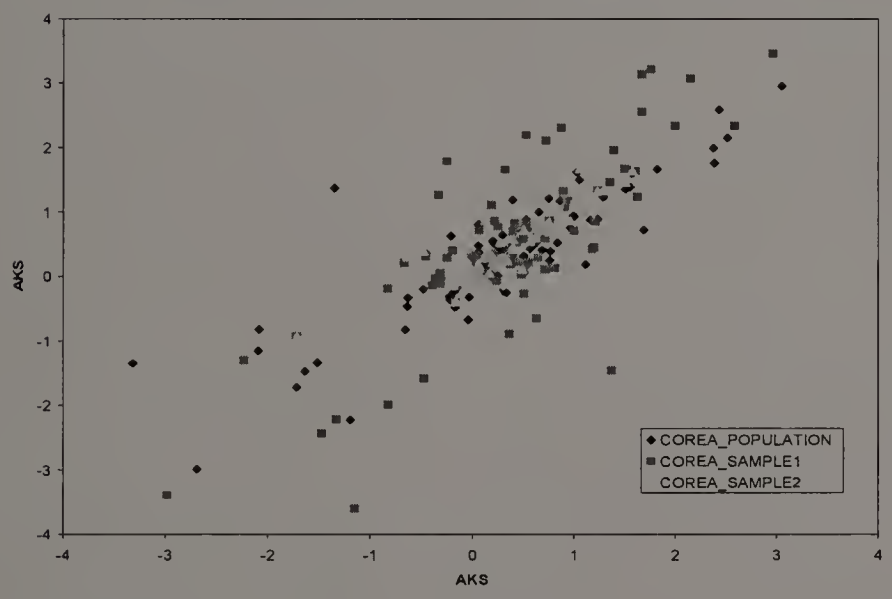
Correlations		Nominal Model			Multiple-choice Model			Polyweighting	
		Population	Sample1	Sample2	Population	Sample1	Sample2	Sample1	Sample2
Nominal Model	Population	1							
	Sample1	.993	1						
	Sample2	.992	.971	1					
Multiple-choice Model	Population	.758	.765	.739	1				
	Sample1	.720	.741	.689	.857	1			
	Sample2	.581	.566	.586	.578	.401*	1		
Poly-weight	Sample1	.953	.958	.934	.676	.666	.597	1	
	Sample2	.960	.942	.965	.679	.642	.617	.974	1

Note. *Low correlation for the two split-half samples within a model.

a)



b)



c)

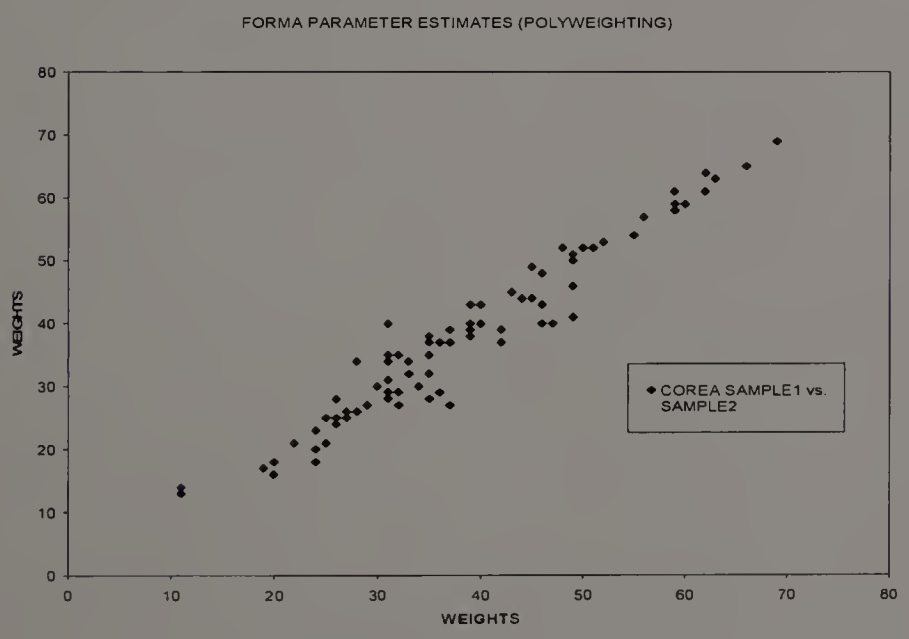
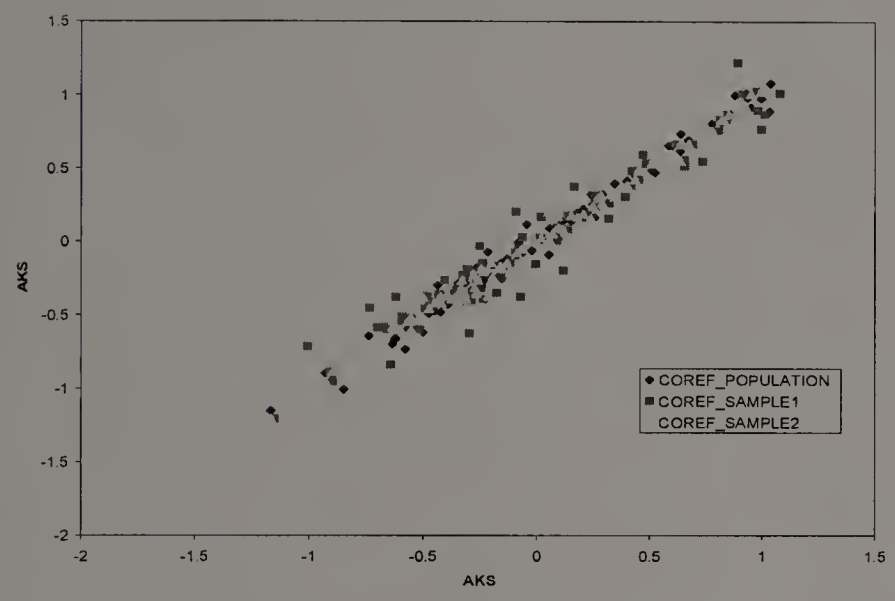
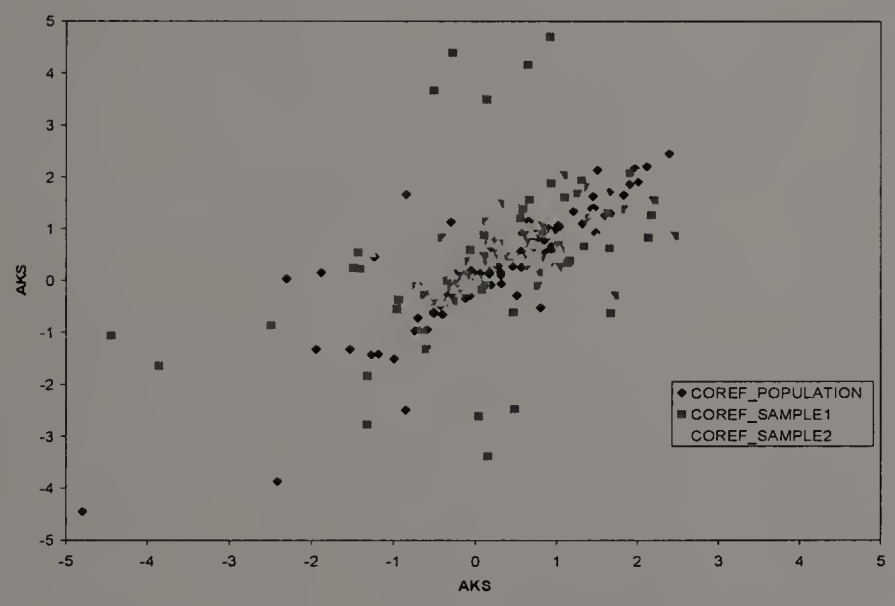


Figure 4.4 Correlations of Option Weights within Models (Form A)

a)



b)



c)

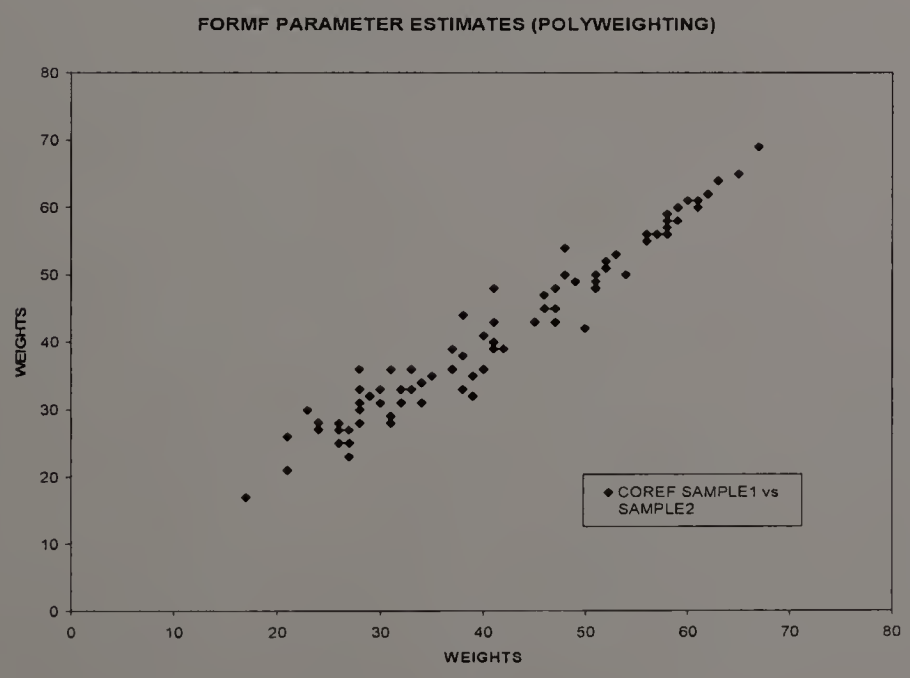


Figure 4.5 Correlations of Option Weights within Models (Form F)

4.4 Summary of the Results

Reliability estimates, theta at the passing score point, and the percent agreement between models and criterion classification for each of the scoring models are summarized in Tables 4.11 and 4.12.

Table 4.11 Summary of the Results (Form A)

Model		Criterion		
		Marginal Reliability	I_θ at the Passing Score (theta = -0.27)	Percent Agreement between Model and Criterion
3PL Model	Population	.90	14.543	0.816
	Sample1	.90	14.909	
	Sample2	.90	15.343	
Nominal Model	Population	.90	15.178	0.814
	Sample1	.91	15.062	
	Sample2	.91	15.892	
Multiple-choice Model	Population	.91	15.894	0.822
	Sample1	.91	16.103	
	Sample2	.91	17.111	
Polyweighting	Sample1	.91*		
	Sample2	.91*		0.933**

Note. *Coefficient alpha. ** Sample2 statistic.

Table 4.12 Summary of the Results (Form F)

Model		Criterion		
		Marginal Reliability	I_θ at the Passing Score (theta = -0.27)	Percent Agreement between Model and Criterion
3PL Model	Population	.90	12.866	0.796
	Sample1	.90	13.654	
	Sample2	.90	13.078	
Nominal Model	Population	.90	13.812	0.695
	Sample1	.90	14.369	
	Sample2	.90	13.862	
Multiple-choice Model	Population	.91	14.109	0.743
	Sample1	.91	15.268	
	Sample2	.91	16.383	
Polyweighting	Sample1	.91*		
	Sample2	.91*		0.938**

Note. *Coefficient alpha. **Sample2 statistic.

CHAPTER 5

SUMMARY AND CONCLUSIONS

5.1 Summary of the Study

Multiple response items closely resemble multiple-choice items in that both are objective measures of an examinee's knowledge and skills in a content area; they differ, however, in item presentation and response mode. Multiple-choice items have one correct answer whereas MR items typically elicit more than one correct answer from the examinees. With this added complexity in item presentation and response mode, MR items have the potential to minimize guessing on the MC tests. Moreover, this type of item provides an avenue to enhance partial knowledge representation on a test. Because of these advantages, this type of item has gained popularity in achievement aptitude, licensure, and certification testing in recent years.

With the increased use of the multiple response items in achievement, aptitude, licensure, and certification tests comes the question of how to score them appropriately in practice. In many testing programs, multiple response (MR) items are conventionally scored dichotomously either at the item level or the option level, with 1 point for each correct answer and no credit for an incorrect answer. In recent years, however, several testing programs that utilize the MR format began to explore partial credit scoring algorithms to score the MR items. Various models have been suggested for polytomous scoring of MR items and the underlying assumption for these models is that proficiency distributions are not the same for examinees who answer different items correctly or who choose different wrong-answer response options, even if their number right scores

are identical. When test items are scored dichotomously, potentially useful information about an individual's level of proficiency that is contained in the complete pattern of the item responses is lost and the precision with which the test measures is reduced. Polytomous scoring attempts to remedy the situation by incorporating information accrued from examinees' responses to items to obtain a more accurate assessment of individual differences in proficiency.

Partial credit scoring of MR items often involves some type of option weighting of the response alternatives of the MR items. A correct response is given a higher scoring weight than an incorrect one and a less "wrong" response option generally has a larger scoring weight associated with it than a wrong option. It is expected that option weighting can extract additional information from a given set of items; and by so doing improve the precision of measurement. Both classical option weighting methods and polytomous IRT models have been proposed and studied in the past two decades to examine the efficacy of these different models in the context of scoring MC (MR) items polytomously.

In this study, four polytomous scoring models were investigated and the effectiveness of each model was evaluated using data from a large-scale certification exam. The four scoring models examined were: polyweighting model, three-parameter logistic model, Bock's nominal model, and the multiple-choice model of Thissen and Steinberg. Of the four models, three were IRT models and one was a classical option-weighting model. Evaluation criteria, which included the reliabilities, item and test information functions, as well as the accuracy of pass-fail decisions, were employed to examine the viability of each of the scoring models.

For the comparison between dichotomous and polytomous scoring methods, the multiple response items were scored dichotomously and calibrated along with the rest of the multiple-choice items using the three-parameter logistic model. The resulting item and test information functions were compared to the information functions obtained under the polytomous IRT models.

For comparison between classical and IRT polytomous scoring, the test was scored by polyweighting procedure. Scoring weights derived from the calibration sample were used to score item responses in the application sample and the resulting scores and the option weights were used for comparisons between linear and IRT polytomous scoring methods.

The results obtained from various scoring models suggested that dichotomous scoring model performed equally well as the polytomous scoring model in regards to overall reliabilities. As discussed in Chapter 4, the reliability estimates remained virtually unchanged across different scoring models. This is in contrast to previous studies which show that weighting of the response options of a MC test generally results in an increase in overall reliability of the test, since more components (i.e., response options) of the test were included in the calculation of the internal consistency reliability estimates. A possible explanation for the unexpected outcome observed in this study is that 2/3 of the items were multiple-choice items and were scored the same way regardless of the scoring methods used. It is also possible that due to the sufficiently high reliabilities obtained with the dichotomous scoring model, scoring MR items polytomously is less likely to produce a significant increase in over reliability of the tests.

Results from this study are line with previous studies with respect to test information functions. Little or no increases were observed at the passing score point and the upper end of the score scale when different scoring models were applied, but polytomous scoring resulted an increase in test information function, and the increase is substantial at the lower end of the score scale. This is conceivable, as pointed out by Thissen (1976), for examinees of higher proficiency are less likely to select incorrect choices and accordingly there is less information available in the incorrect responses for the upper half of the proficiency range. It follows that the more difficult the test is, the more incorrect responses may be expected; and the more incorrect responses that are available, the more improvement may be expected from polytomous scoring. Other studies bearing on the merits of polytomous scoring reached similar conclusions (Drasgow, Levine, Williams, McLaughlin, & Candell, 1989; Levine & Drasgow, 1983; Sympson, 1983, 1993;Thissen & Steinberg, 1984).

Since information function is the inverse of the standard error of measurement, an increase in information simultaneously translates into a decrease in measurement error. As the standard error of measurement decreases, the accuracy with which the ability is estimated is improved. In other words, polytomous scoring improves the precision of ability estimation for the lower half of the proficiency range. These findings are in agreement with previous research in which polytomous scoring is shown to generally produce more reliable ability estimates for examinees, particular for low ability examinees (Bock, 1972; Thissen & Steinberg, 1984; Thissen, Steinberg, & Mooney, 1989).

With respect to classification accuracy, studies have shown that empirical option weighting typically produced slightly more reliable domain score estimates and more consistent pass-fail decisions than number-right scoring, particularly in the lower half of the test score distribution (Haladyna, 1990; Sympson & Haladyna, 1988). Huynh and Casteel (1987) evaluated the usefulness of Bock's nominal response model with respect to the validity of pass-fail decisions. They found that the use of Bock's nominal model for moderate-length tests did not produce decisions that differ substantially from those based on raw scores and the validity of those decisions did not change noticeably when different ability estimates were used (i.e., raw scores vs. Bock ability estimates). They did observe, however, that when the test was short, the pass-fail decisions based on Bock ability estimates and those based on raw scores were in less agreement for examinees at the lower end of the ability range. Findings from this study are inconsistent with previous research, especially in the cases of IRT model-based polytomous scoring. The discrepancy rates between the model and criterion classification, as evaluated by using dichotomous scoring of MC items as criterion classification, were inconsistent across models when polytomous scoring was applied. This moderately high discrepancy rate, when translates into the number of examinees that were being misclassified, poses a serious threat to the validity of the test score interpretation and use in certification testing situations.

In summary, the results from this study do not suggest an advantage of polytomous scoring over the dichotomous scoring with respect to overall measurement precision (i.e., total score reliability and information at the passing score point). However, polytomous scoring did improve the precision of measurement at specific

score points (e.g., the lower end of the score scale). The increase in information at the lower end of the score scale is impressive given only 20 multiple response items were rescored using different polytomous scoring methods.

Although the increases in test score reliabilities and information functions observed in this study were minor, the practical importance of the increases has much to recommend the use of option weighting in some cases. It has been recognized that option weighting can extract additional information from a given set of items. This allows the test developer to use fewer items in a test, while retaining a previously set reliability standard. This, in turn, is especially desirable in the case where items are difficult and/or expensive to construct.

This feature is especially useful in computerized adaptive testing (CAT) since many of the item selection algorithms implemented in CAT use maximum information approach, which directs the next best item administered to be the one that provides the most information at the examinee's current ability estimate, based on the responses to the previous items administered. The addition of information extracted from incorrect responses enables the item selection algorithms to reach the desired level of precision with fewer items in a shorter period of time. Studies done by De Ayala (1989, 1992) comparing the efficacy of nominal model and three-parameter model scoring in a CAT environment found that while the two models performed equally well, considerably fewer items were administered by the nominal model CAT than the three-parameter logistic model CAT. This is because the nominal response model provides more information than the three-parameter logistic model for examinees of low proficiencies.

It is clear from the foregoing discussion that polytomous scoring generally increase the precision of ability estimation, particularly over the lower half of the proficiency range, by using information accrued in incorrect responses. The extent to which the benefits of polytomous scoring can be realized largely depends on the characteristics of the items and the examinee population, For MR items of moderate difficulty, polytomous scoring could be more effective than dichotomous scoring because IRT polytomous scoring can capitalize on the information contained in incorrect answers. For a test that contains easy MR items, polytomous scoring may not be very appropriate because less information would be available in incorrect responses, consequently, these items have less power to differentiate among examinees of varying ability levels therefore the benefit of having a sophisticated scoring system can hardly be realized.

Last but not least important is the appropriate use of polytomous scoring models. Although various polytomous scoring models have been proposed for scoring multi-category data, special attention must be paid to the assumptions and limitations of each of the scoring models when applying them to real MR data. Results from this study suggest that among the polytomous scoring models investigated, the nominal response model appears to be more stable than the multiple-choice model in terms of the accuracy of item and proficiency parameter estimates. The multiple-choice model has an additional parameter (d_k for the DK category) and requires a lot more cycles to reach convergence. In fact, the multiple-choice model rarely converges in many senses because it is usually overparameterized. The best solution to the convergence problem is to give it a lot of cycles or stop it somewhere (personal communication with D.

Thissen through Scientific Software Incorporation, July 23, 2003). In light of this limitation, multiple-choice model is not the ideal model for scoring multi-category data. Hence, nominal response model should be used in scoring MR items when appropriate.

In the comparison of classical versus IRT polytomous scoring models, results suggest that polyweighting would perform equally well as the nominal response model scoring in terms of measurement efficiency, and classification accuracy. Moreover, Polyweighting method is easy to use, does not have strict sample size requirements or the assumptions of unidimensionality that is required in item response models. Thus, when the sample size is small and/or when the data are not unidimensional, polyweighting model should be preferred over the IRT polytomous scoring models.

5.2 Directions for Future Research

The above conclusions are based on the analyses conducted in this study. Clearly further research is warranted. Additional study can be carried out to examine the quality of the MR items and its impact on the information function of the items.

1. The validity of the option weights can be examined by using an external criterion, which the current study clearly does not have. Judges can be brought in to evaluate each of the response options of the MR items and judgmental weights can be determined and assigned. Empirical weights can be compared to judgmental weights to examine the congruence of the weighting, which can also be used as a viable means to validate the judgmental weights.

2. Since MR items are frequently nested within a testlet, research on scoring methods for testlet-based MR items should be useful in examining the quality of the

MR items and in evaluating the variability of polytomous scoring approaches in the context of scoring testlet MR items.

3. The results in the current study are limited by the fact that the test has relatively few MR items. For a test with a higher proportion of MR items, polytomous scoring may yield greater gains in reliability and test information functions. Hence, future research can examine tests with a larger number of MR items to determine whether the gains resulted from polytomous scoring are statistically significant.

4. Using a sample that is larger than the currently studied one to examine the effects different scoring schemes may have on the efficiency of measurement and the accuracy of classification. Findings from the present study suggest that the parameter estimates obtained from the population are more stable compared to the those obtained from samples, both in the nominal model and the multiple-choice model calibration cases, presumably because the item calibration algorithms require a larger number of examinees for accurate estimation of item and person parameters. In the case of polyweighting, the restriction on sample sizes is not as stringent as in IRT calibration; still, a large sample size ensures a more precise estimation of the option weights.

5. Apply MTF scoring to the MR items and compare the number right scoring with MTF scoring. MTF scoring was excluded from the current study because preliminary analysis indicated a poor model fit to the data. In previous studies, individual True-False items were collapsed into a singular item for which a component score can be obtained by summing up the item score of the individual True-False items. In this study, however, each item needs to be stretched into a number of individual items by adding 0s to the categories that were not the keyed response category, resulting

in a large number of 0 responses in most of the newly-generated items. The item calibration software used in this study (i.e., Multilog) failed to produce reasonable parameter estimates for these items. Preliminary results in this study notwithstanding, the comparison is of practical importance as MTF scoring is being used in several testing programs to score MR items. Thus, given model data fit, MTF scoring in the context of MR tests should be studied and results from the comparative analysis can help testing program identify the right scoring method for MR items on a test.

5.3 Conclusions

The appropriateness of the scoring options for MR items largely depends on the set of data and the extent to which the model fits the data. The main findings of this study are:

1. Theoretically, polytomous scoring methods provide a better way to assess examinees' level of achievement, ability, or knowledge. However, these methods should be used with caution when applying to real data.
2. Large sample sizes are preferred in IRT model-based polytomous scoring conditions because of the prerequisite of the sample sizes in IRT model calibrations.
3. Polytomous scoring increases the information functions at the lower end of the score scale, thus, for test with cutoff scores in that region, which many of the licensure and certification tests have, polytomous scoring methods may be more appropriate.

4. Polytomos scoring holds the potential for providing useful diagnostic information about an examinee's level of achievement at a specific score point, therefore, using of the polytomous scoring may enhance the interpretation of test scores and the meaningful use of them in admission, selection, and certification processes.

In summary, classical option weighting and polytomous IRT models appear to improve the psychometric properties of the test scores in specific situations (e.g., increase in overall test information function, and increase in information function at specific score points). Since polytomous scoring of MR items enables the measure of an examinee's partial knowledge, it will likely be more useful for providing diagnostic information for examinees whose scores are at the lower end of the proficiency scale. With this in mind, it is the author's contention that polytomous scoring methods should be applied to score MR items for optimal results when appropriate. The current study represents an attempt to determine the extent to which these scoring methods can be generalized to MR scoring. The results of the study may have some bearing on the issue of appropriate scoring rules for MR items.

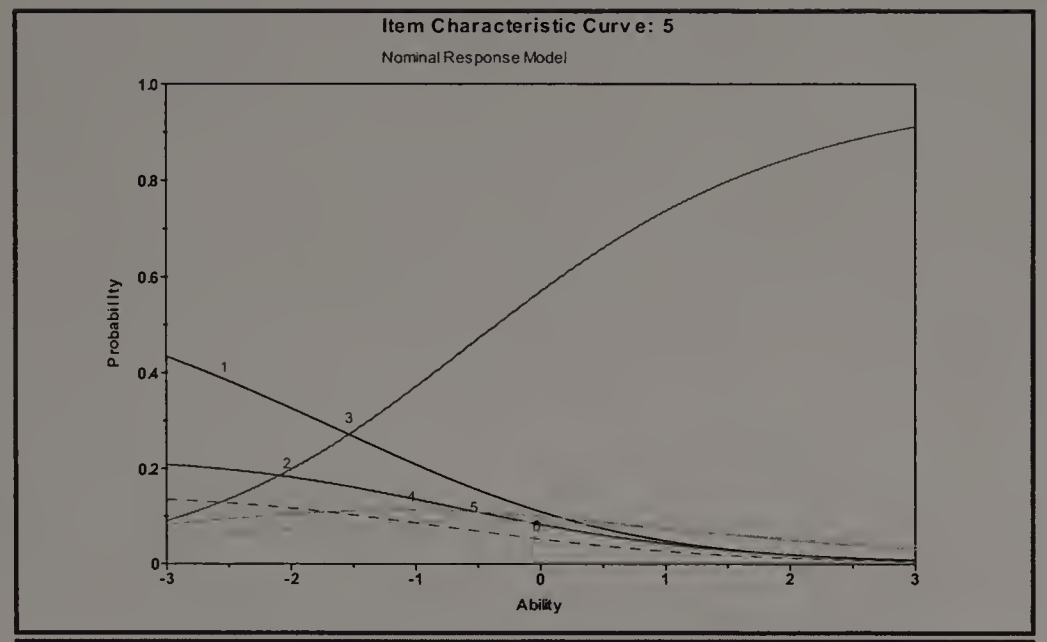
The measures with which we test student's achievement, ability, and aptitude are constantly evolving and so is the environment in which we test. Multiple-choice tests have long been the most frequently used objective measure of students' learning outcomes, but as we move toward computerized testing, there is need and demand to explore other innovative item formats that can measure examinees' true level of knowledge and proficiency in a more authentic way. Multiple-response type of items, because of its flexibility in test construction and scoring, has much to recommend its

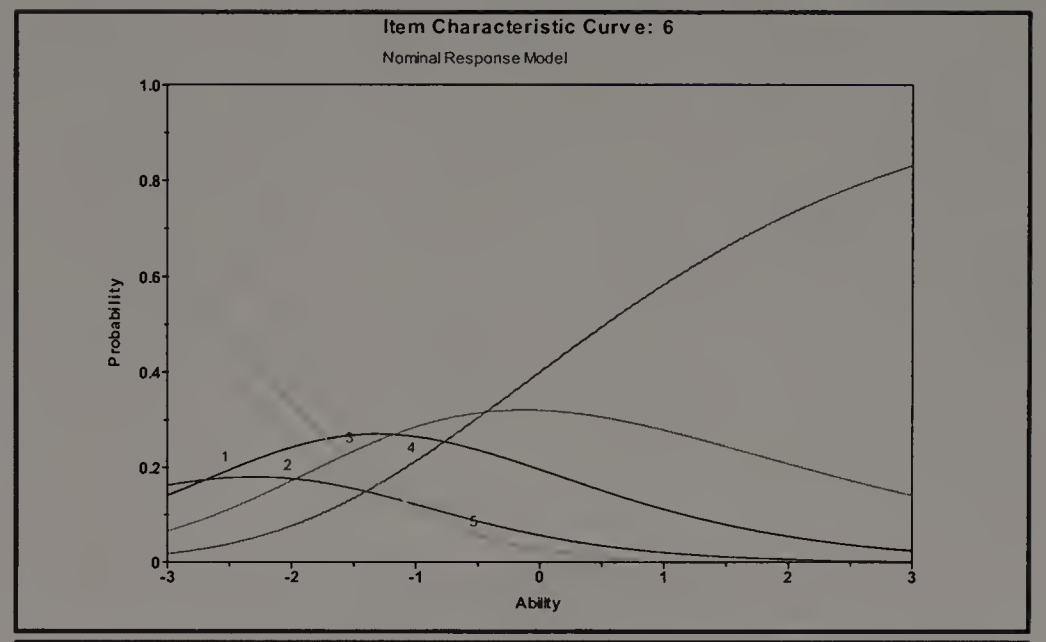
use in achievement and credential testing situations. To realize the benefits afforded by MR items (i.e., partial knowledge representation, finer discrimination among examinees), scoring approaches other than the conventionally used dichotomous scoring method need to be investigated and their advantages and disadvantages documented. The current study examined the MR item format and the associated polytomous scoring approaches in an attempt to provide empirical guidance to the scoring of MR item format in operational testing. For testing programs that use MR items on their tests, findings from this study can be of practical importance to them. It is obvious from this study that polytomous scoring is preferred for tests with fewer items and for tests where precise ability estimates are required like in the credentialing testing. These empirical guidelines can help testing programs determine what the best approach it is to score MR items.

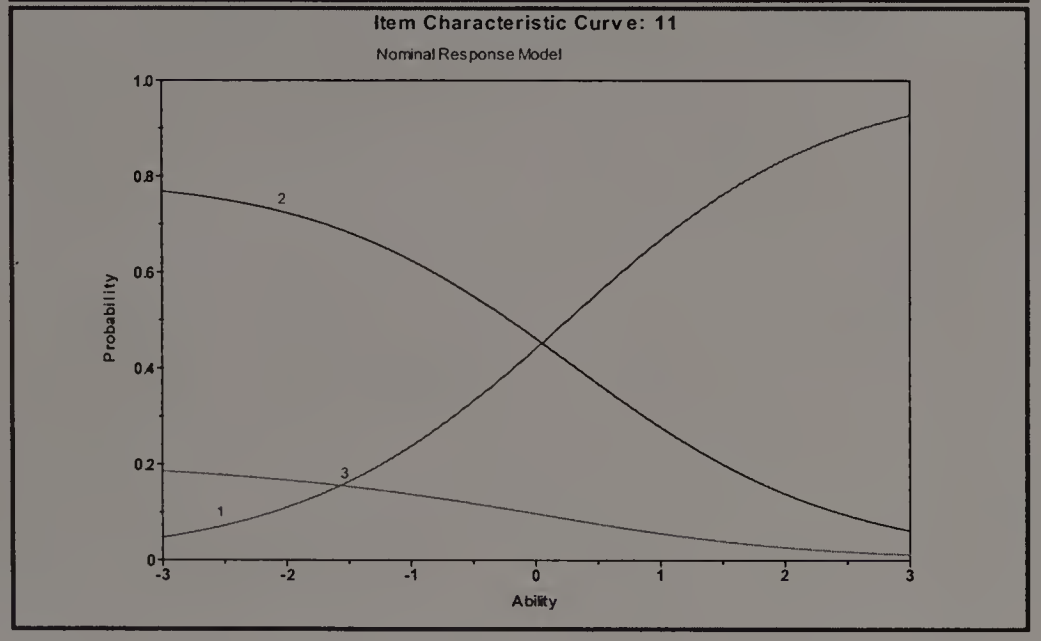
The present study is only the beginning of a continuous effort to investigate innovative item formats and alternative scoring approaches. More research is needed to further advance our understanding of the effectiveness of different scoring approaches. It is expected that studies on this item format and related scoring approaches will receive more scholarly attention in the future.

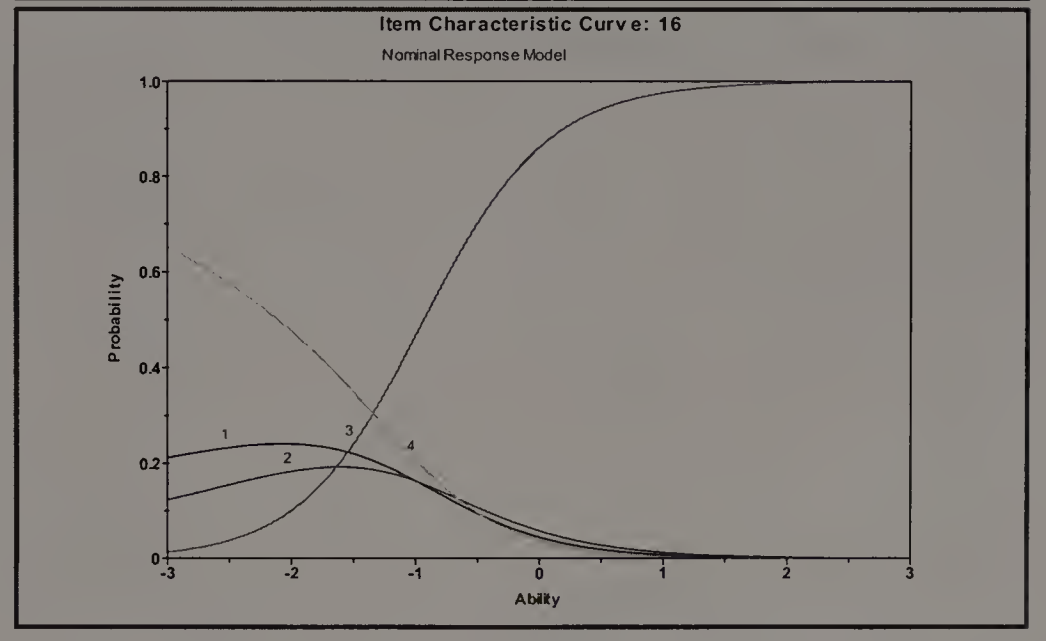
APPENDIX A

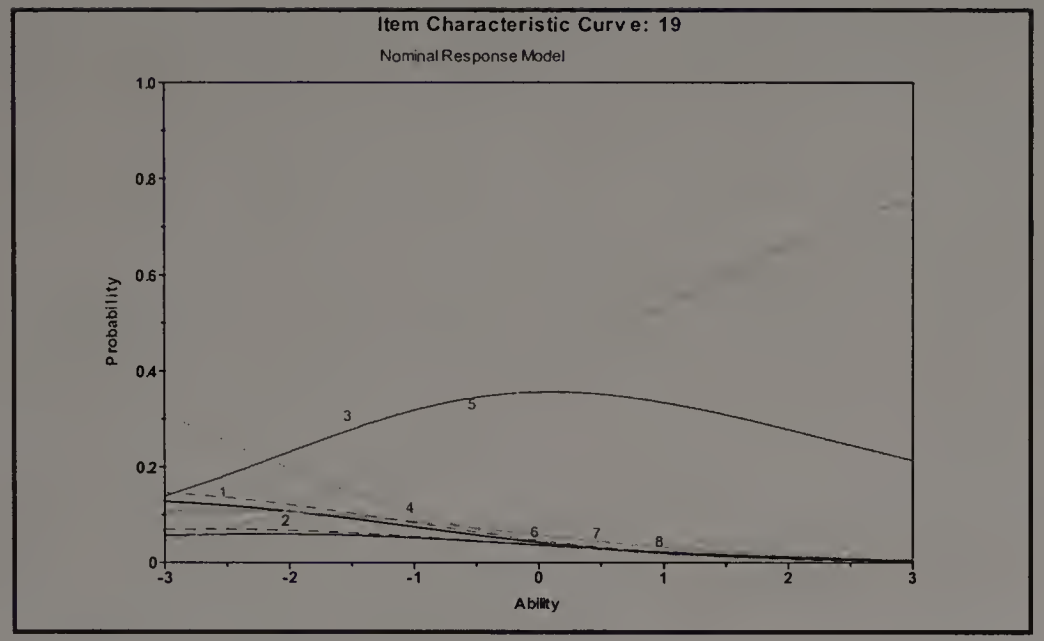
FORM A POPULATION ITEM CATEGORY RESPONSE FUNCTIONS (NOMINAL RESPONSE MODEL)

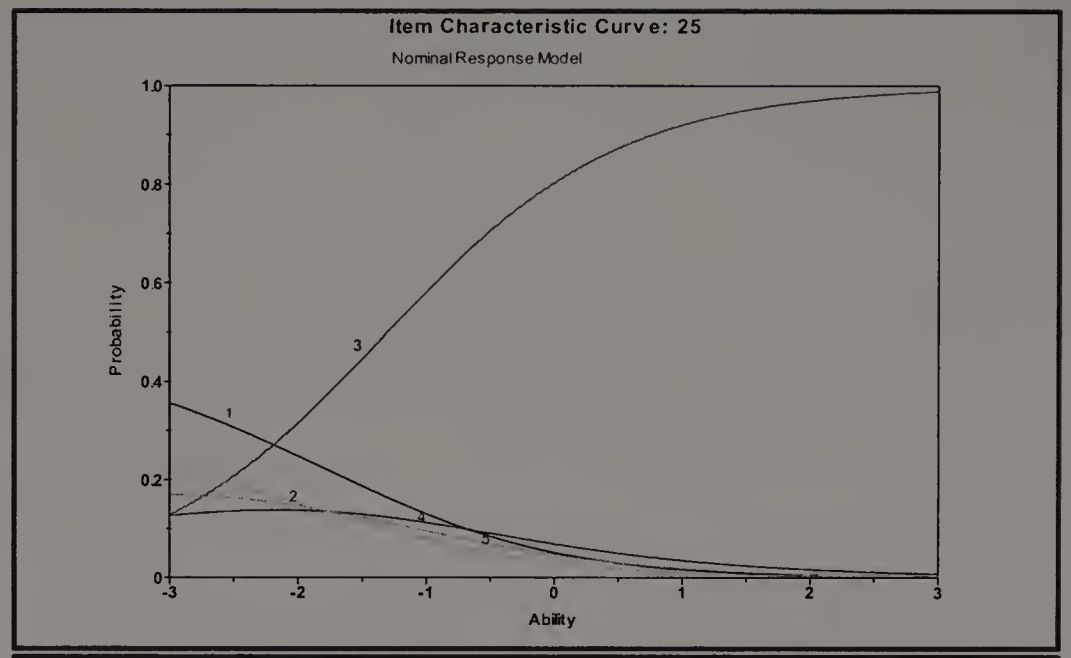

Item Characteristic Curve: 5
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5
6

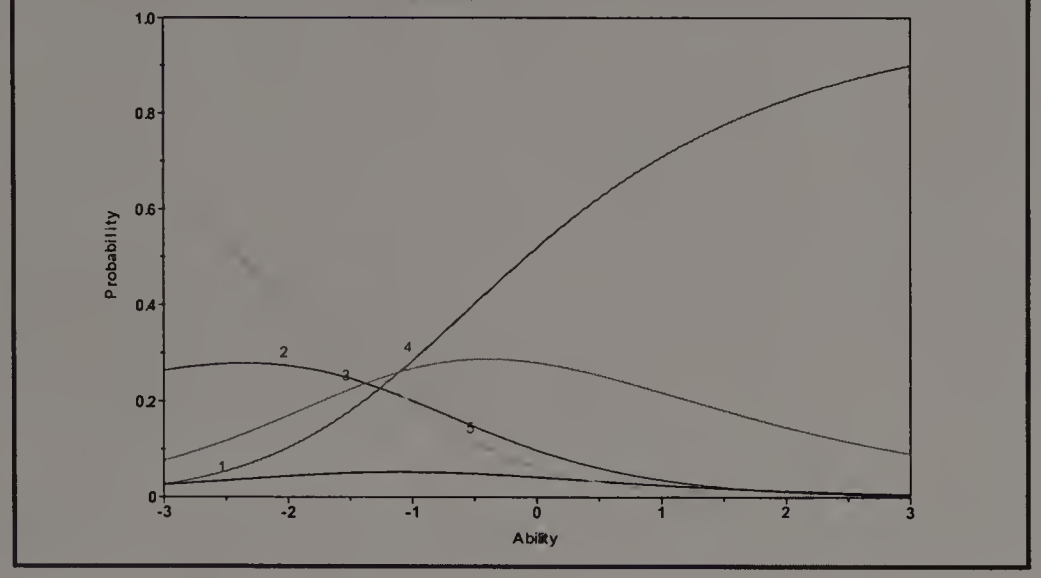

Item Characteristic Curve: 6
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5

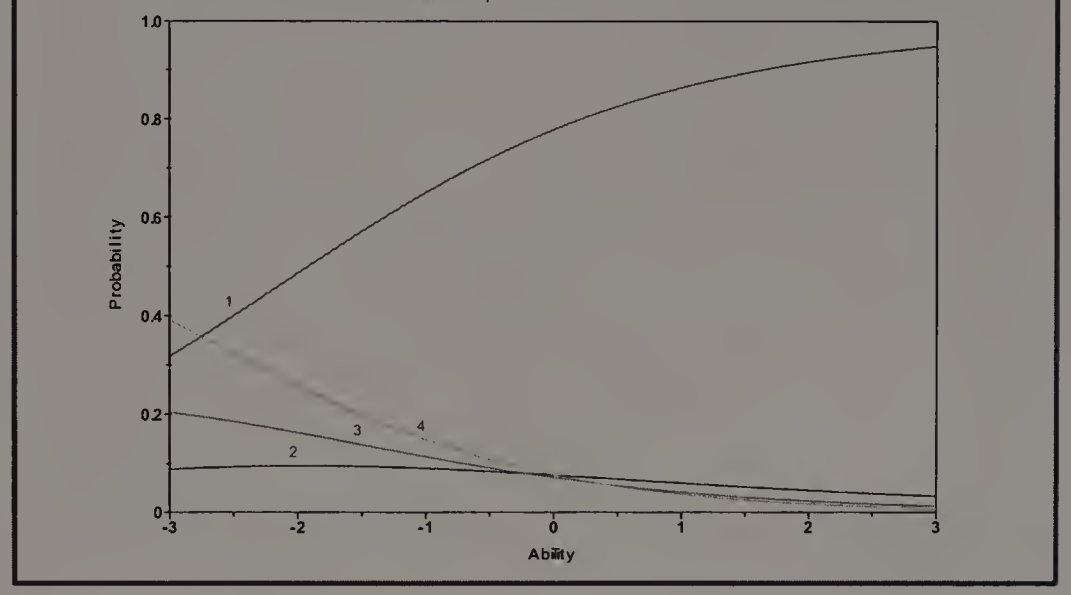

Item Characteristic Curve: 11
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3

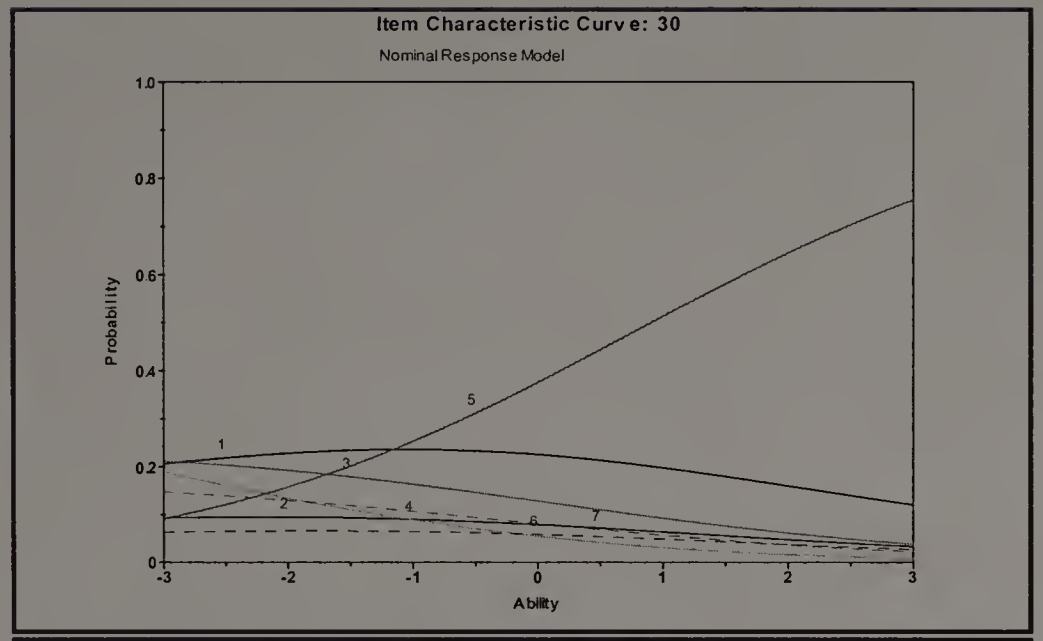

Item Characteristic Curve: 16
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4

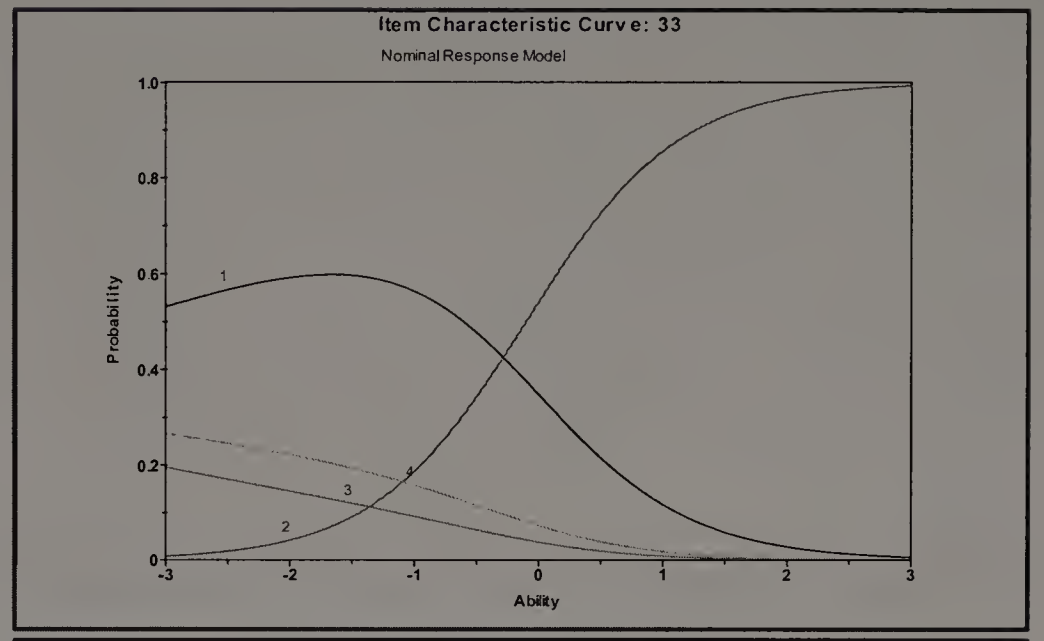

Item Characteristic Curve: 19
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5
6
7
8

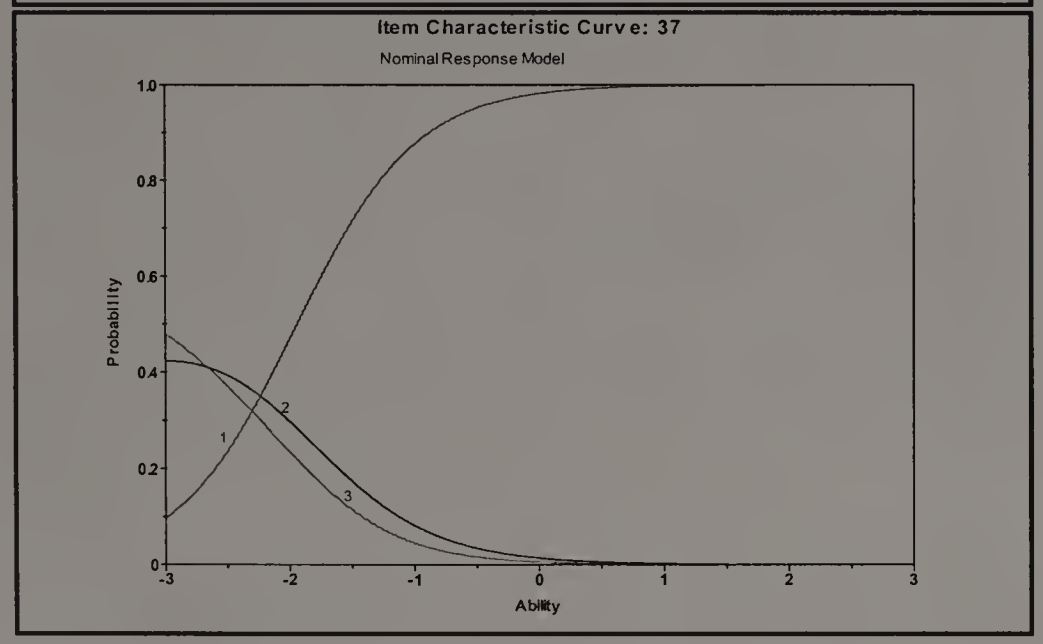

Item Characteristic Curve: 25
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5

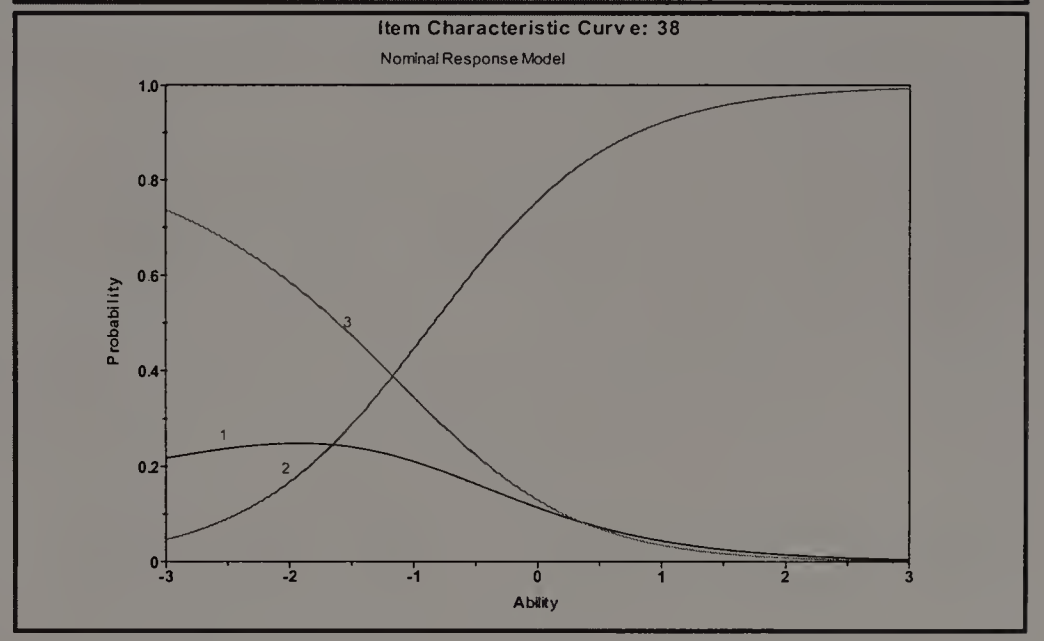

Item Characteristic Curve: 27
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5

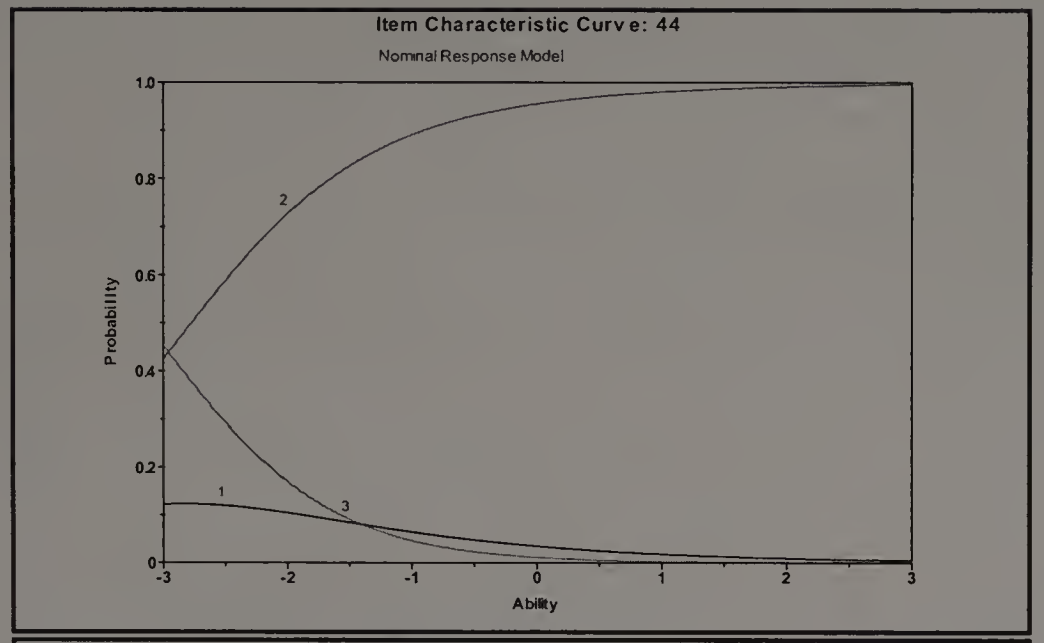

Item Characteristic Curve: 28
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4

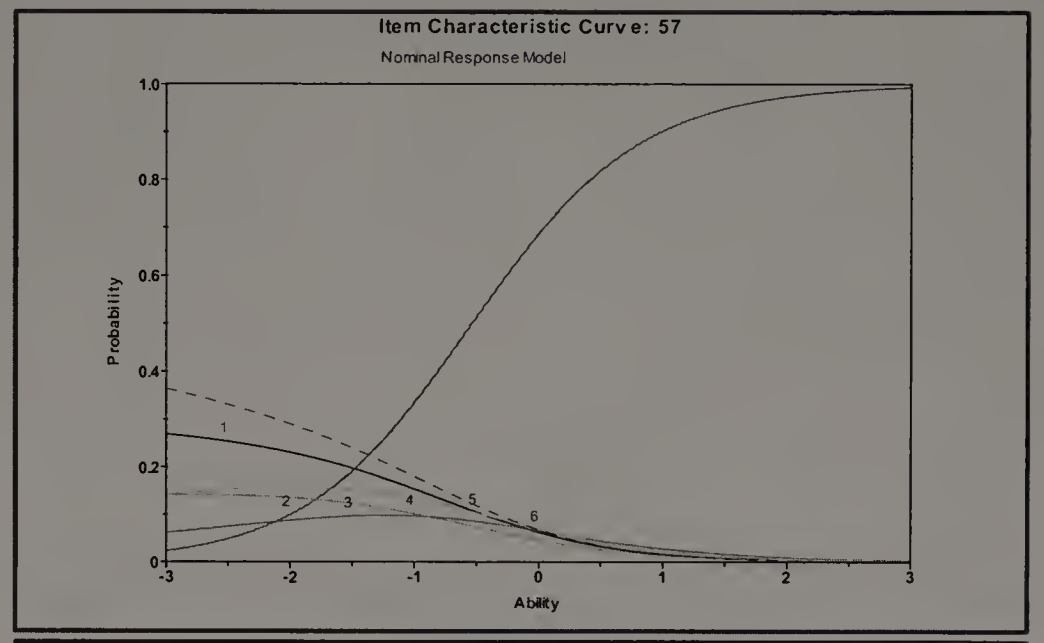

Item Characteristic Curve: 30
Nominal Response Model
Probability
Ability
1
2
3
4
5
6
7

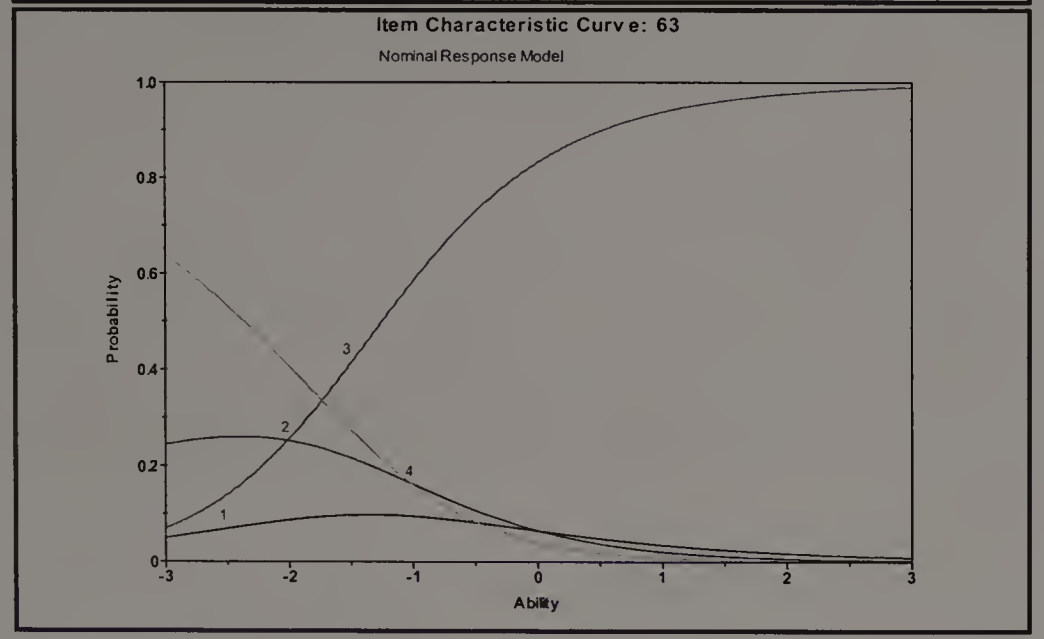

Item Characteristic Curve: 33
Nominal Response Model
Probability
Ability
1
2
3
4

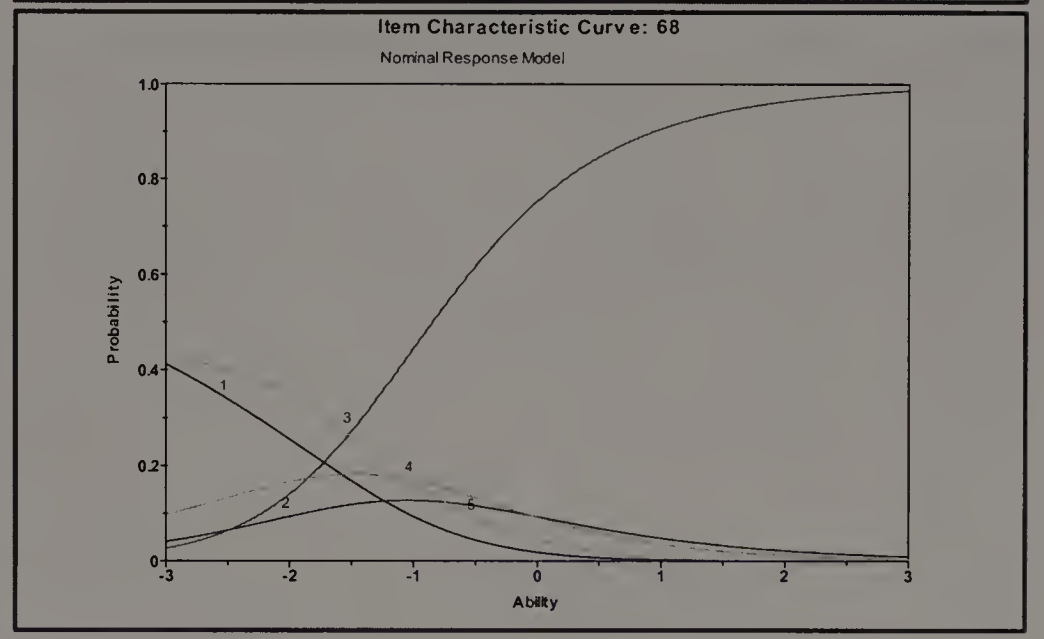

Item Characteristic Curve: 37
Nominal Response Model
Probability
Ability
1
2
3

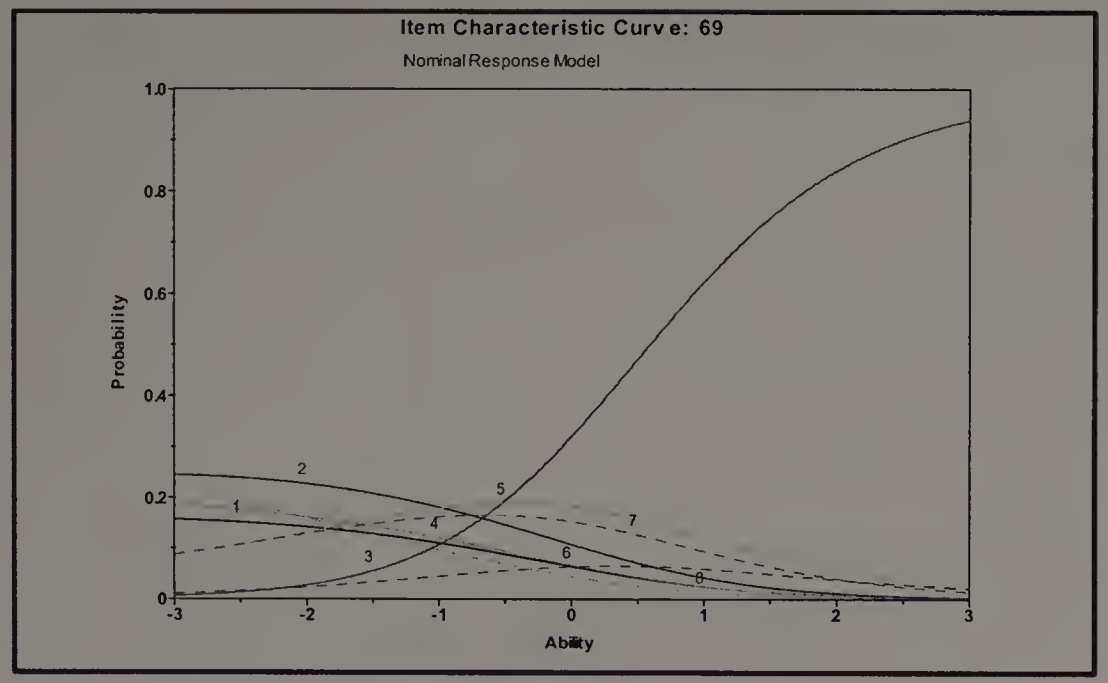

Item Characteristic Curve: 38
Nominal Response Model
Probability
Ability
1
2
3


Item Characteristic Curve: 44
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3


Item Characteristic Curve: 57
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5
6

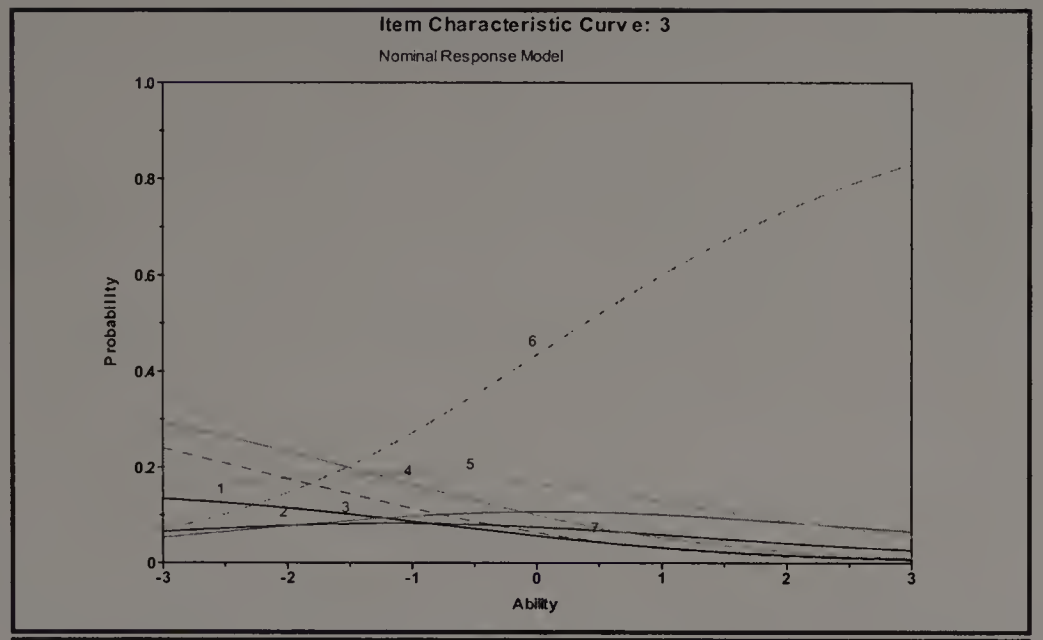

Item Characteristic Curve: 63
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4

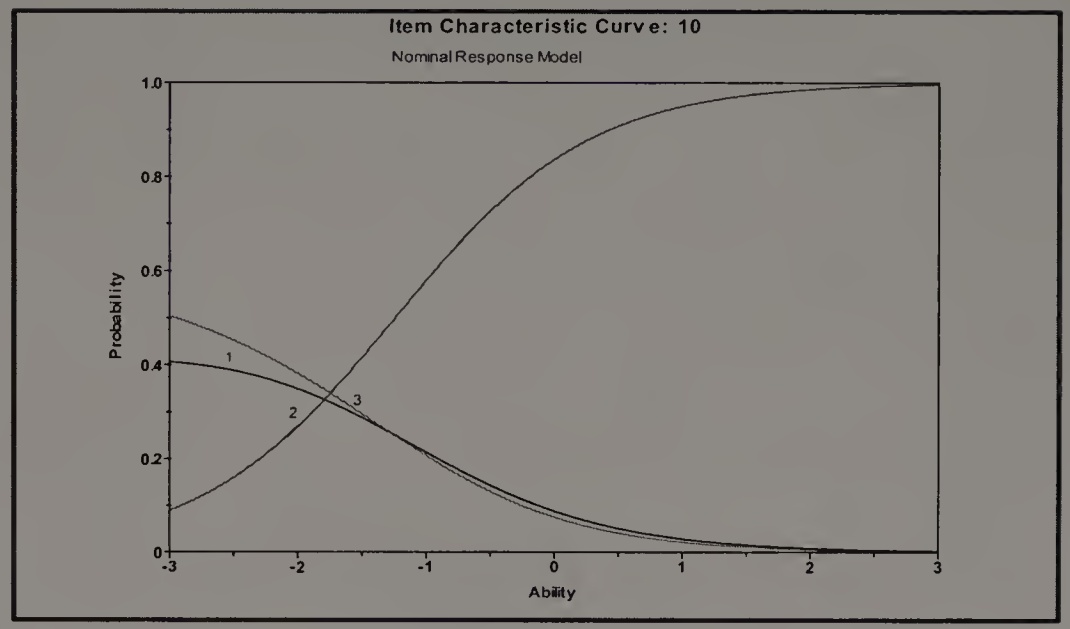

Item Characteristic Curve: 68
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5

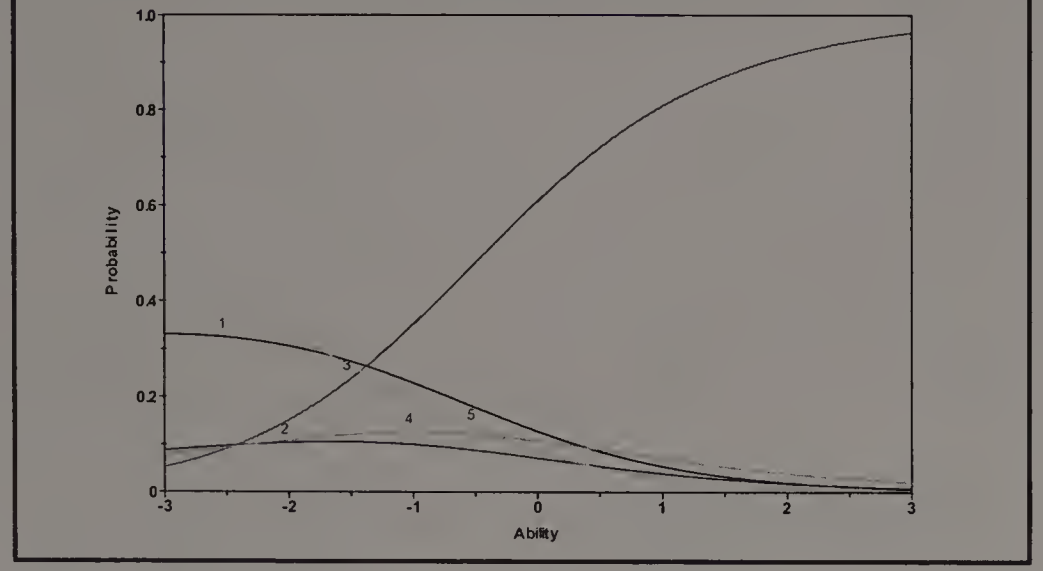

Item Characteristic Curve: 69
Nominal Response Model
Probability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
Ability
1
2
3
4
5
6
7
8

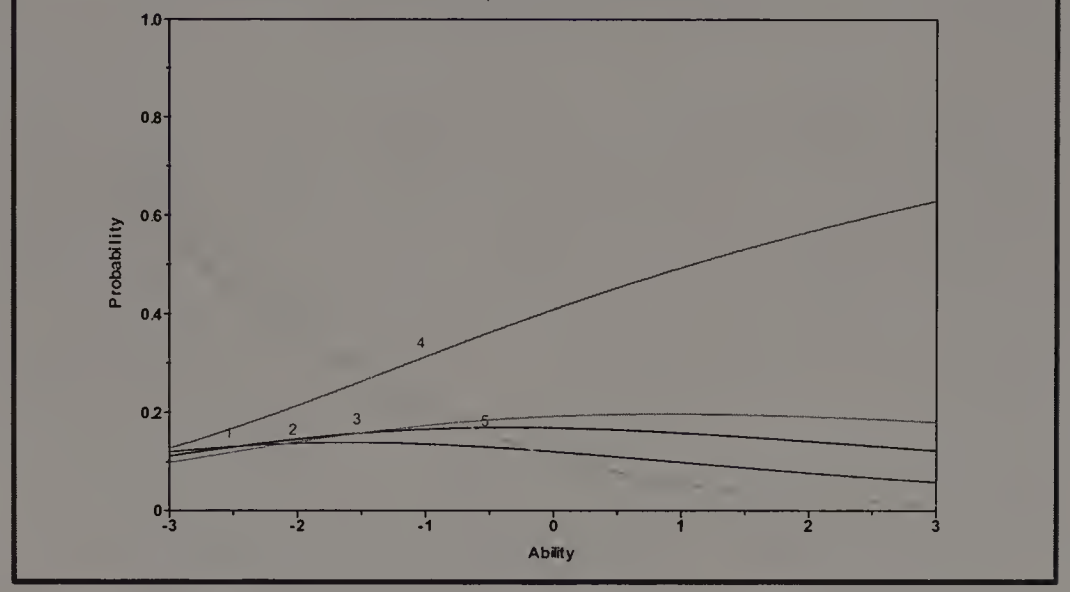
APPENDIX B

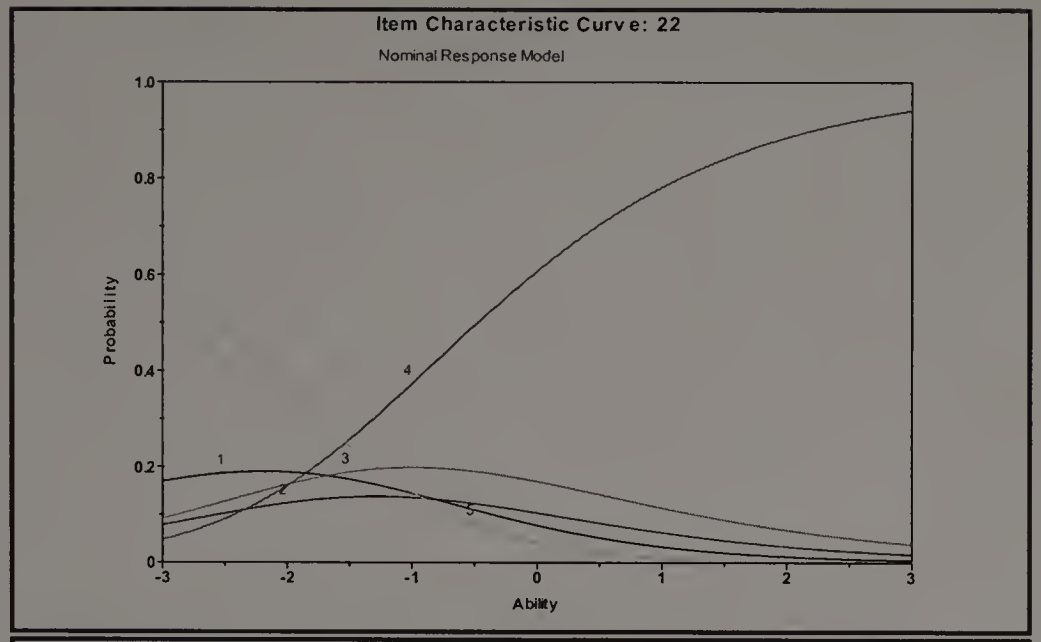
FORM F POPULATION ITEM CATEGORY RESPONSE FUNCTIONS (NOMINAL RESPONSE MODEL)

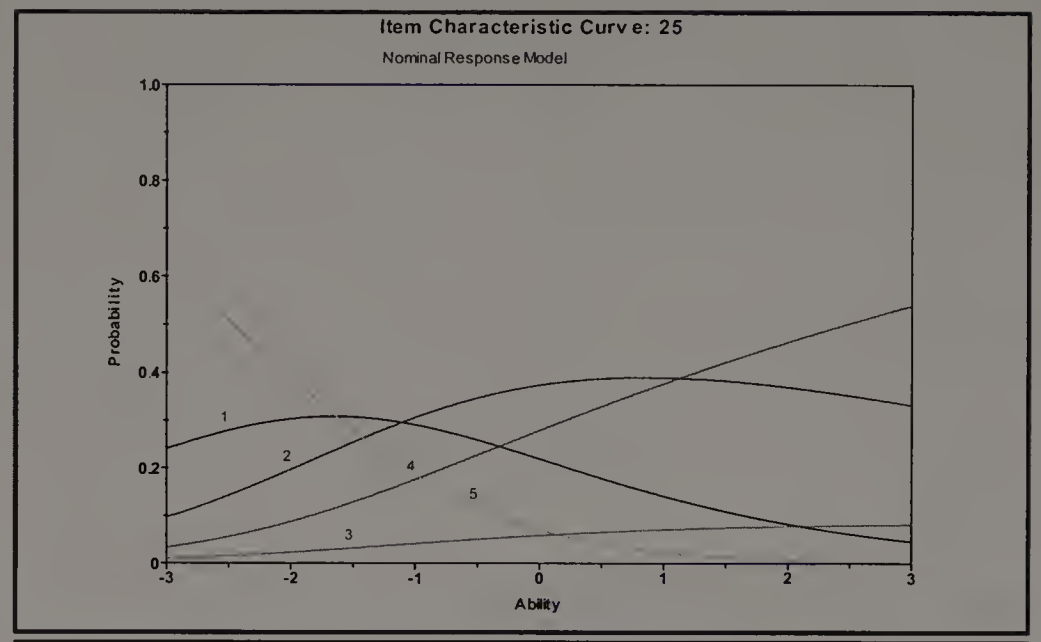

Item Characteristic Curve: 3
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5
6
7

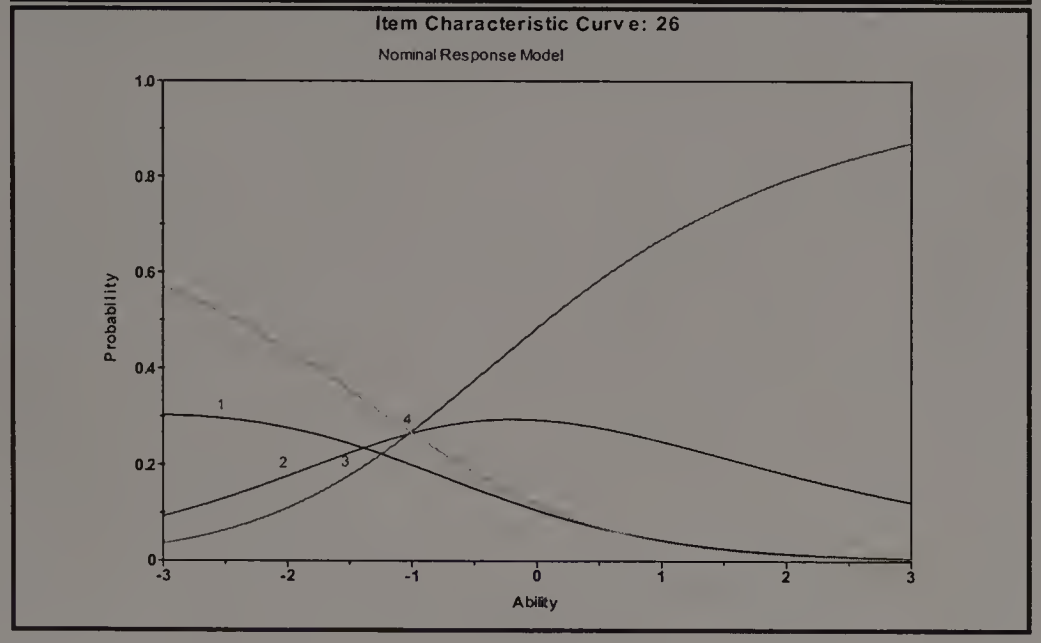

Item Characteristic Curve: 10
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3

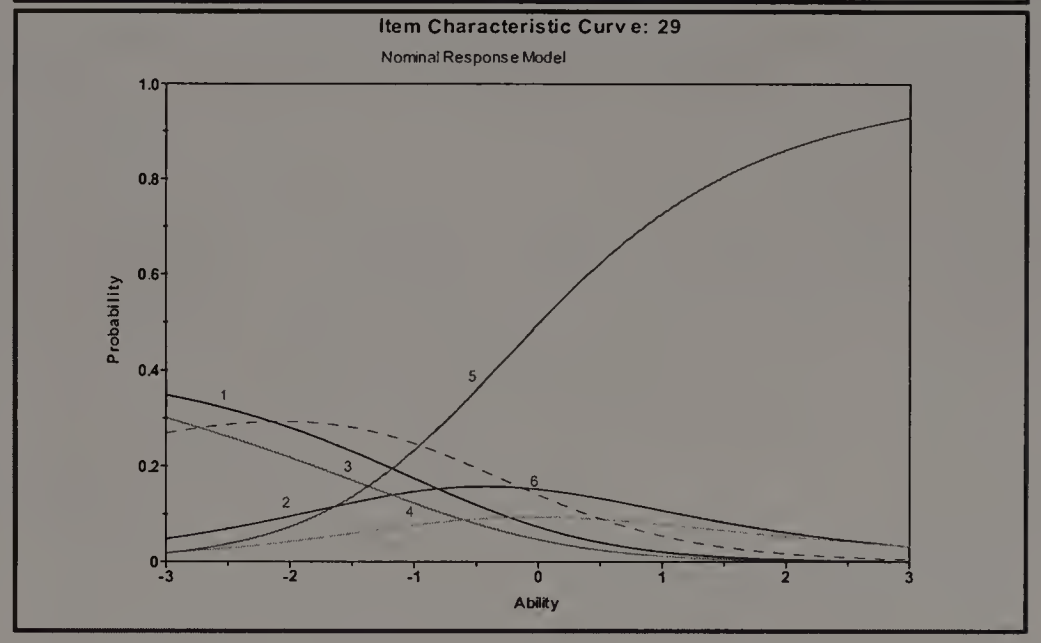

Item Characteristic Curve: 16
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5

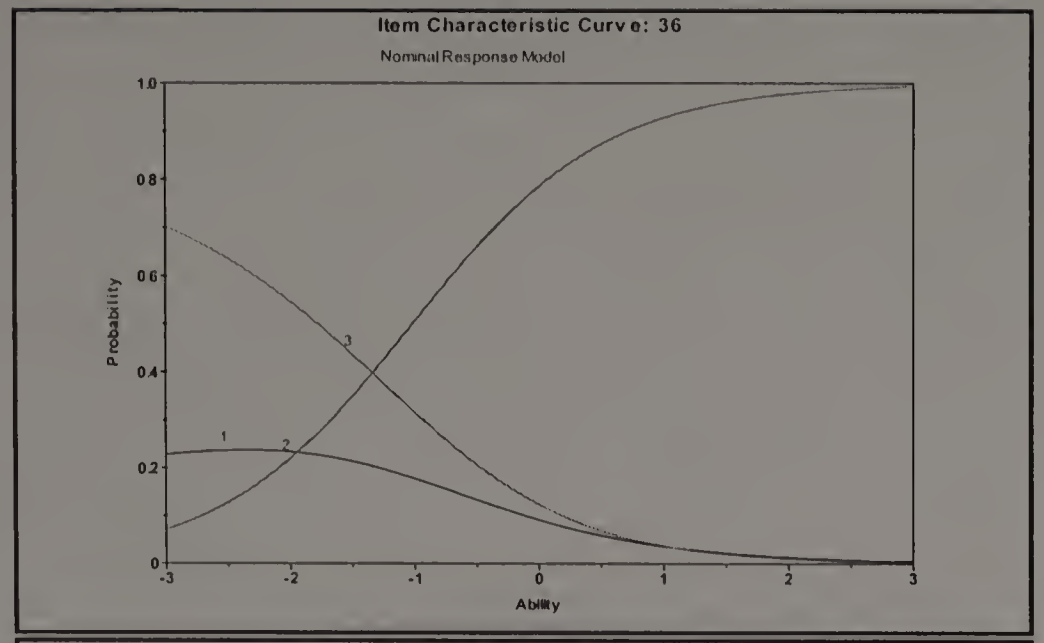

Item Characteristic Curve: 21
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5

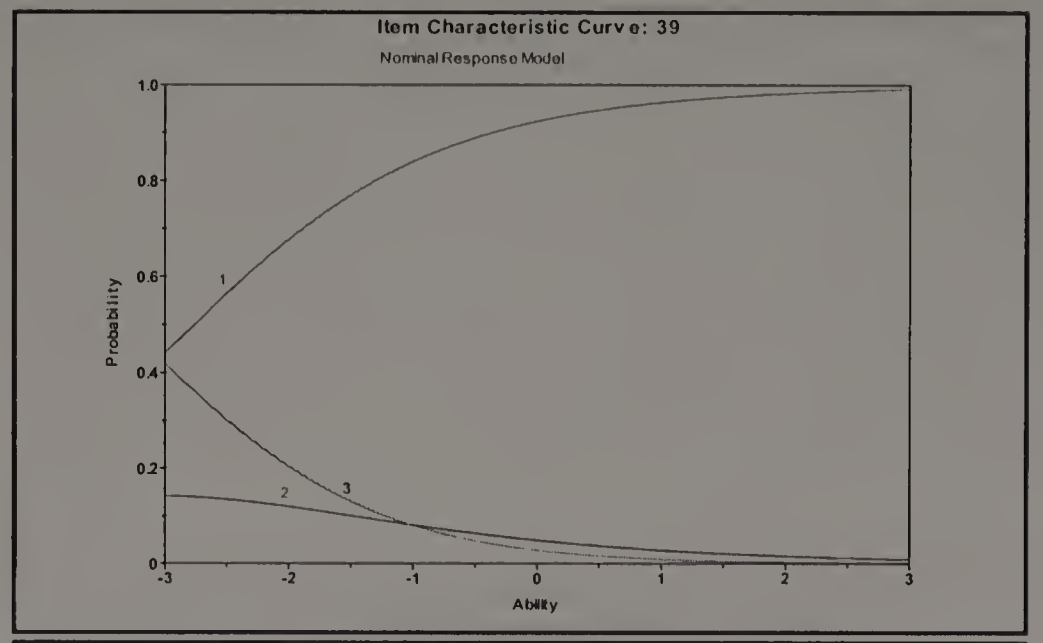

Item Characteristic Curve: 22
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5

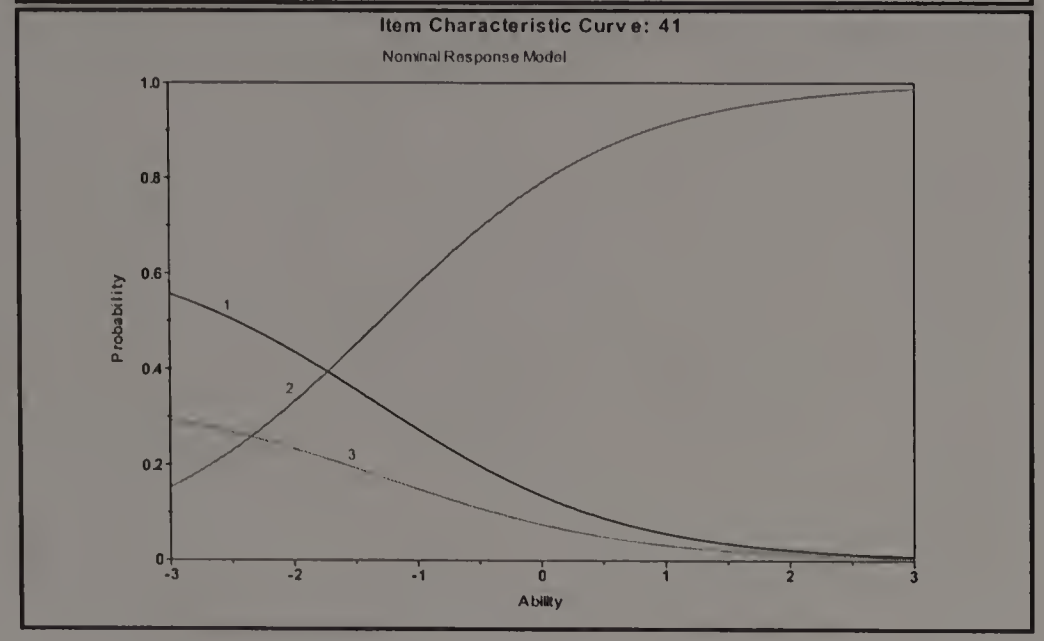

Item Characteristic Curve: 25
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5

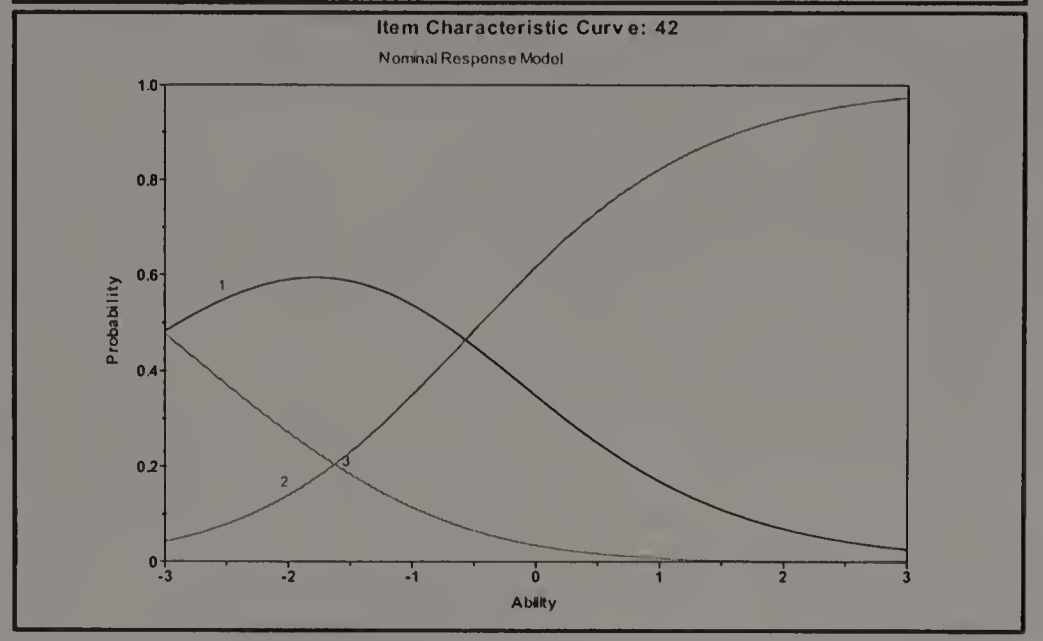

Item Characteristic Curve: 26
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4

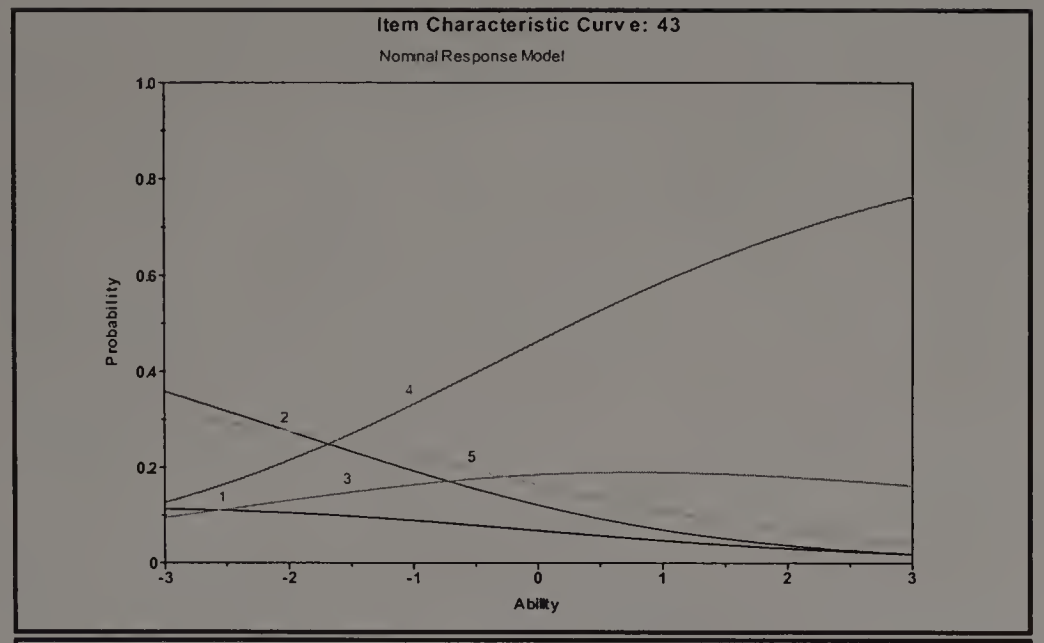

Item Characteristic Curve: 29
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5
6

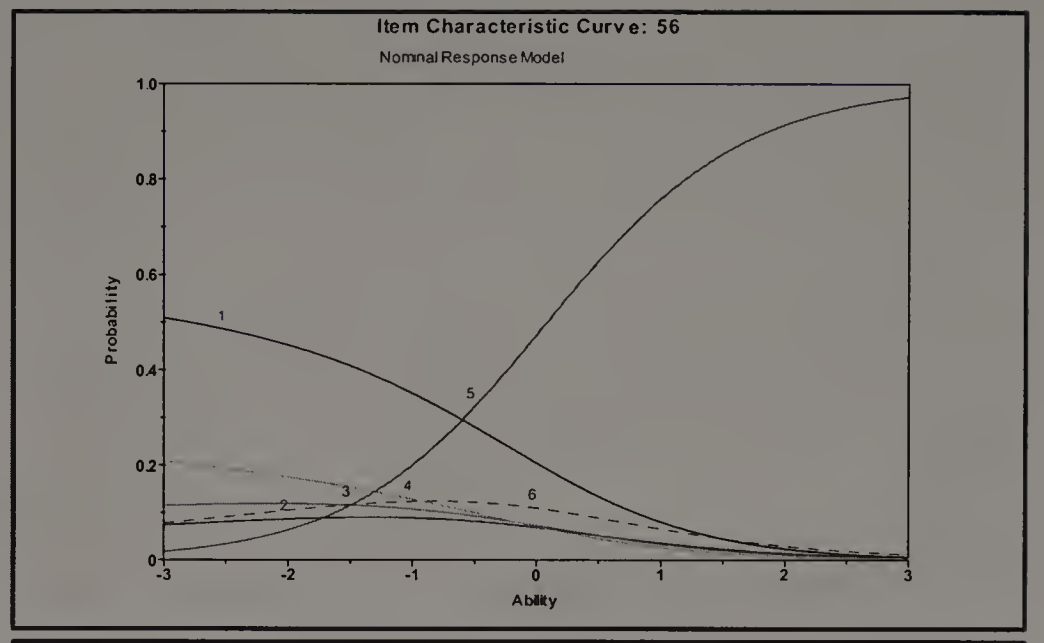

Item Characteristic Curve: 36
Nominal Response Model
Probability
Ability
1
2
3

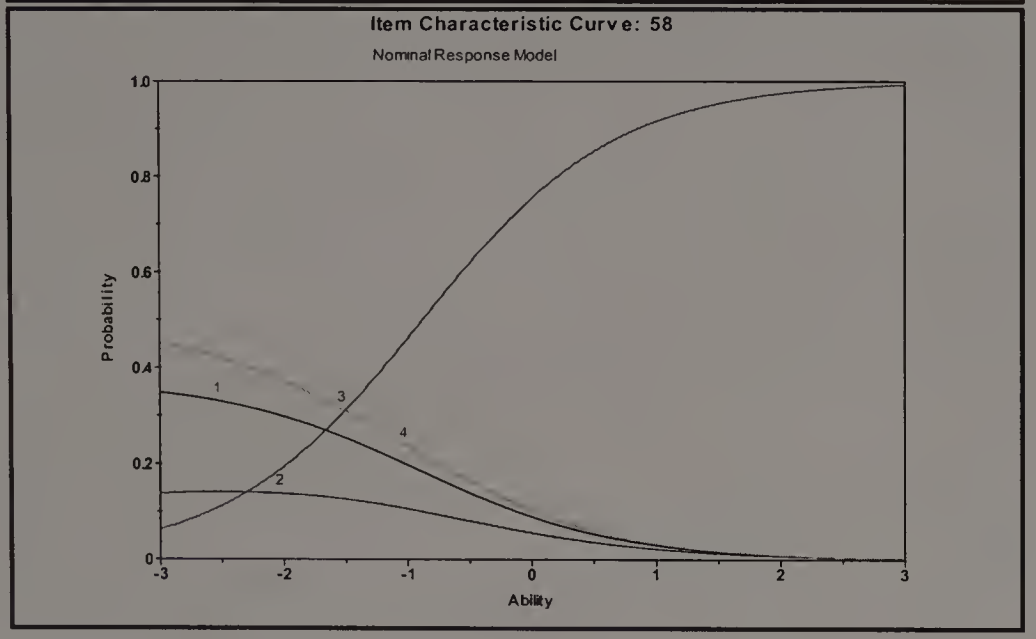

Item Characteristic Curve: 39
Nominal Response Model
Probability
Ability
1
2
3

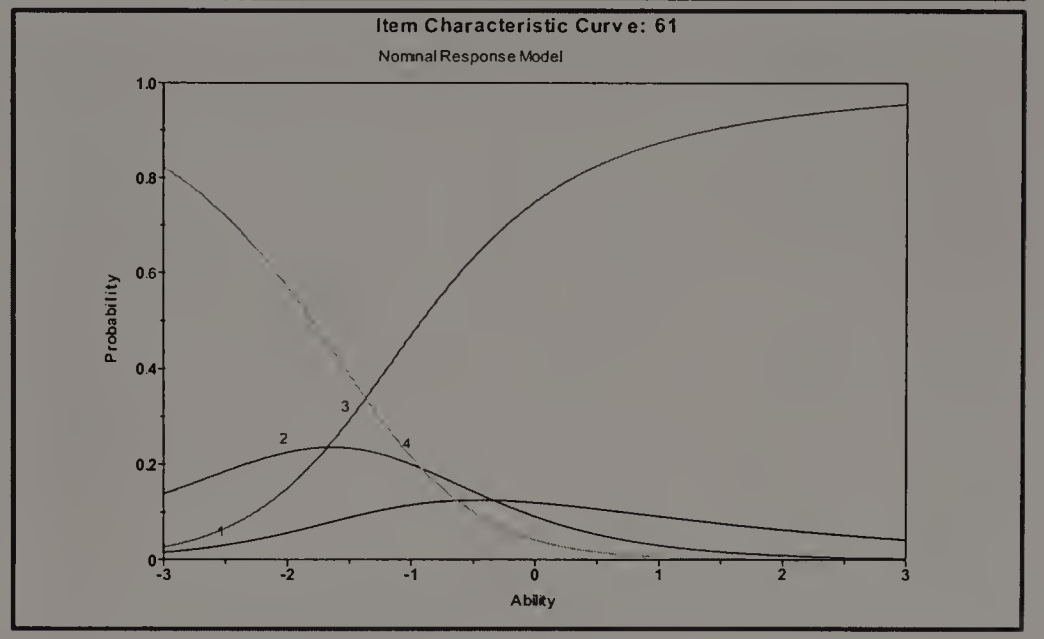

Item Characteristic Curve: 41
Nominal Response Model
Probability
Ability
1
2
3

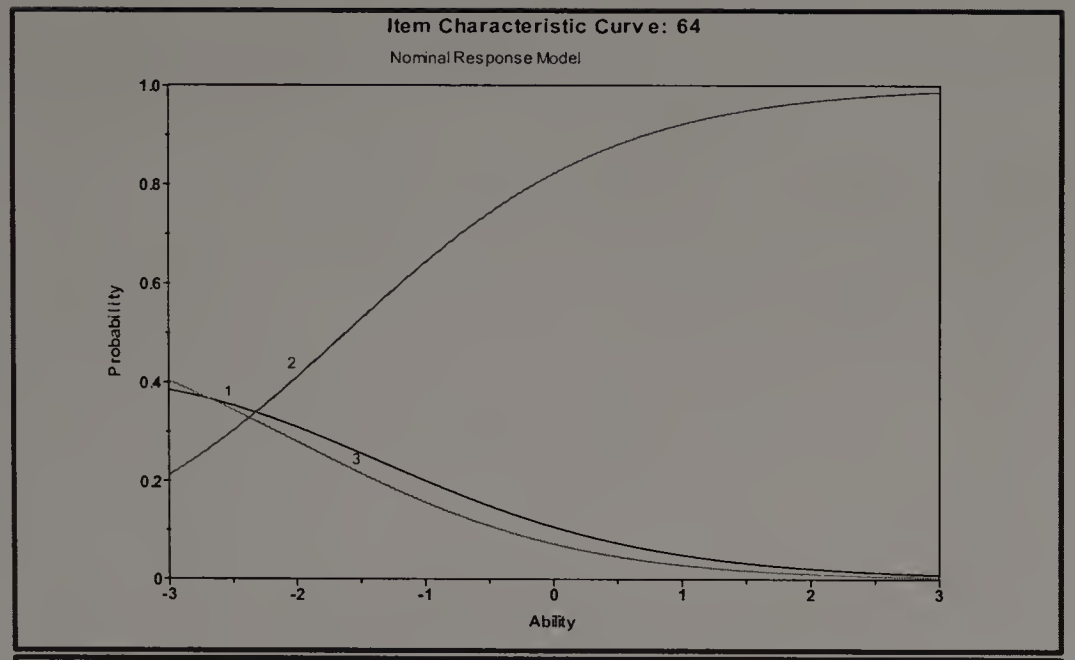

Item Characteristic Curve: 42
Nominal Response Model
Probability
Ability
1
2
3

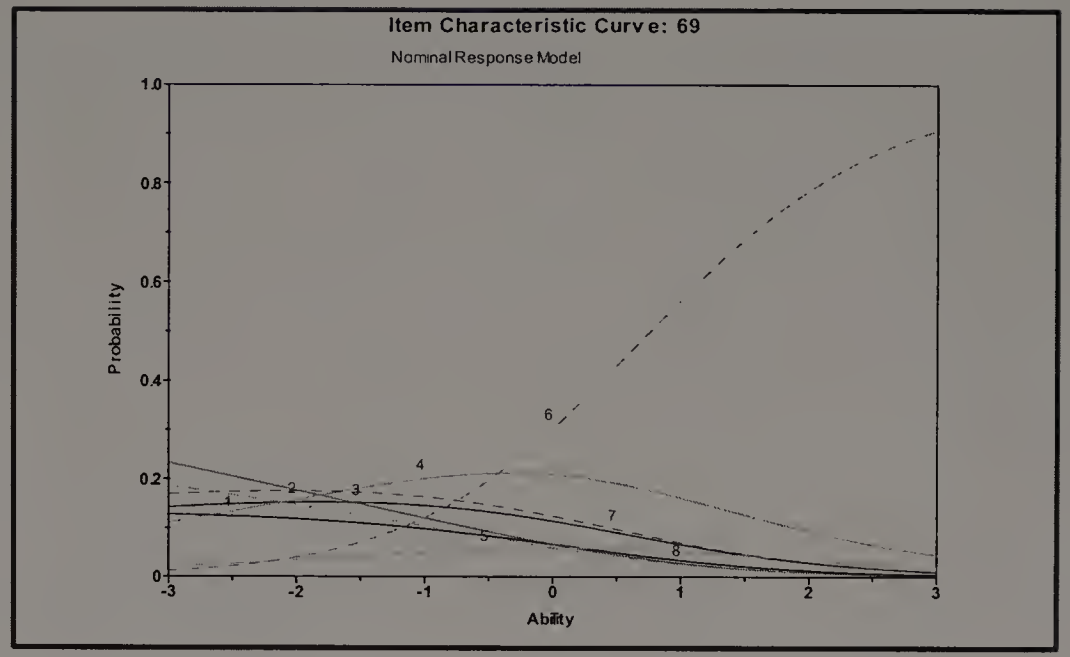

Item Characteristic Curve: 43
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5

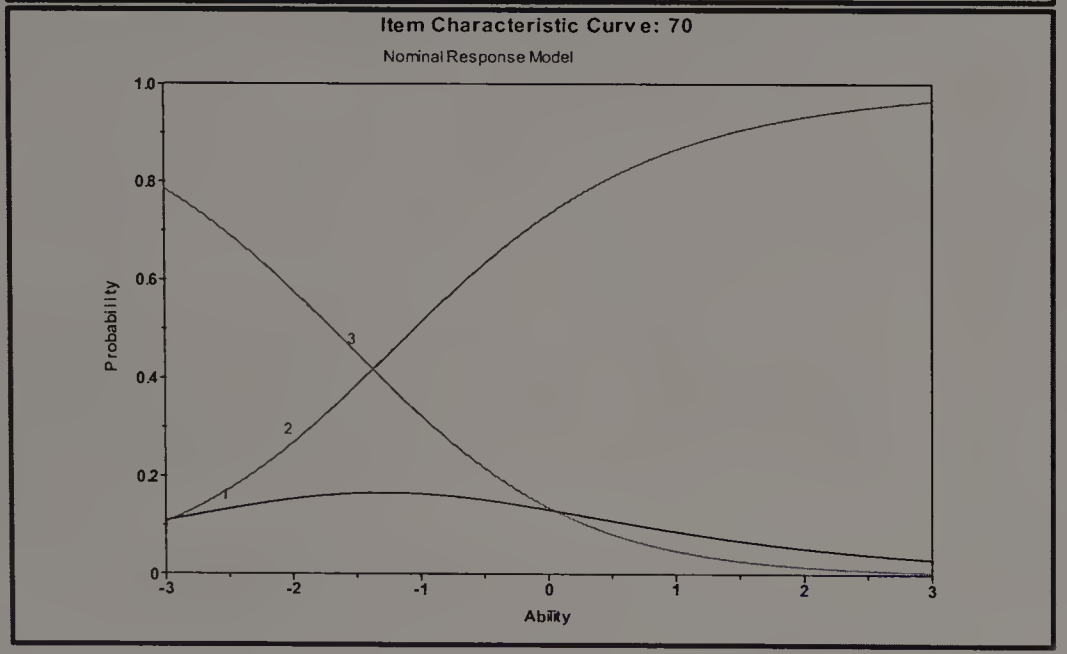

Item Characteristic Curve: 56
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5
6


Item Characteristic Curve: 58
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4


Item Characteristic Curve: 61
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4

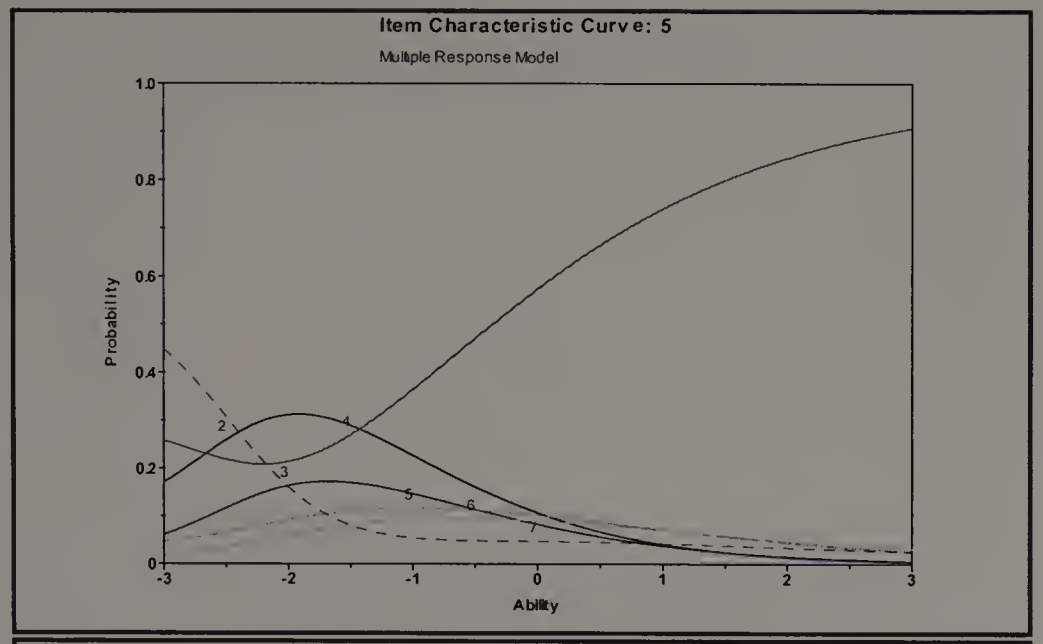

Item Characteristic Curve: 64
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3

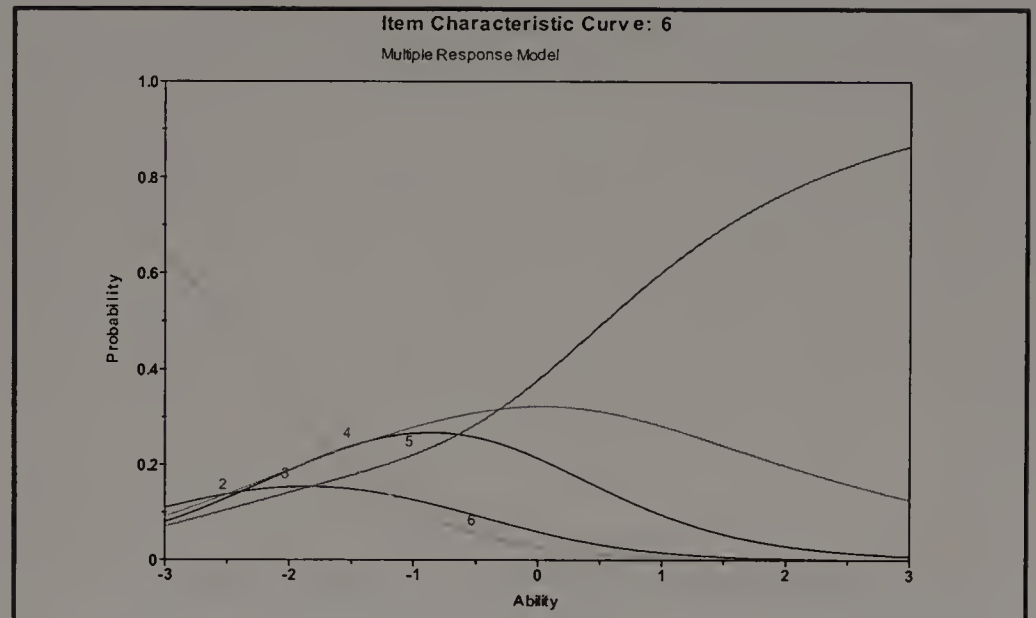

Item Characteristic Curve: 69
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5
6
7
8

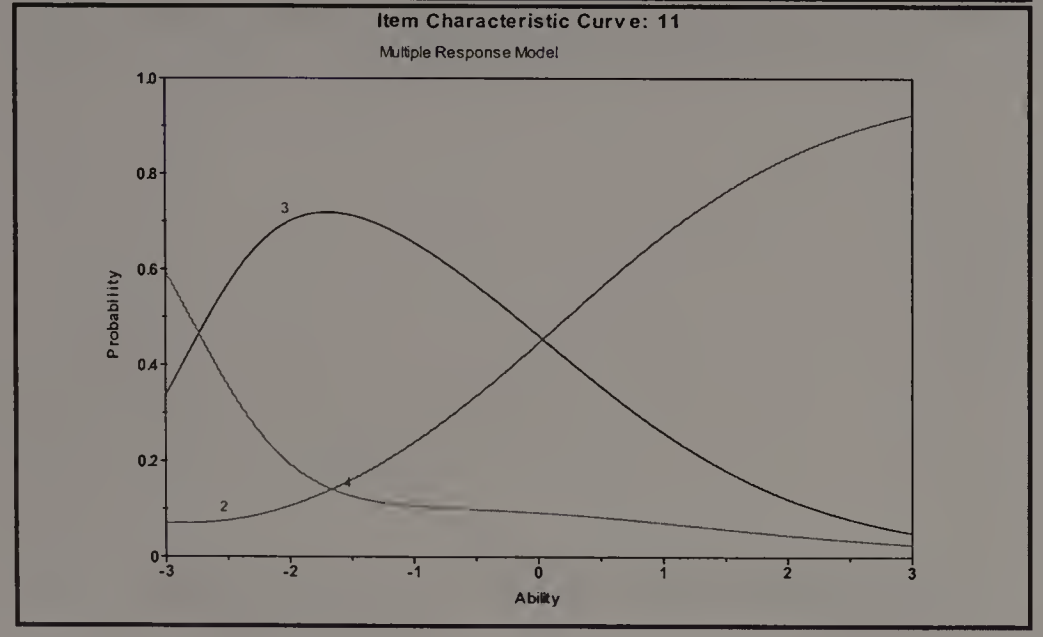

Item Characteristic Curve: 70
Nominal Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3

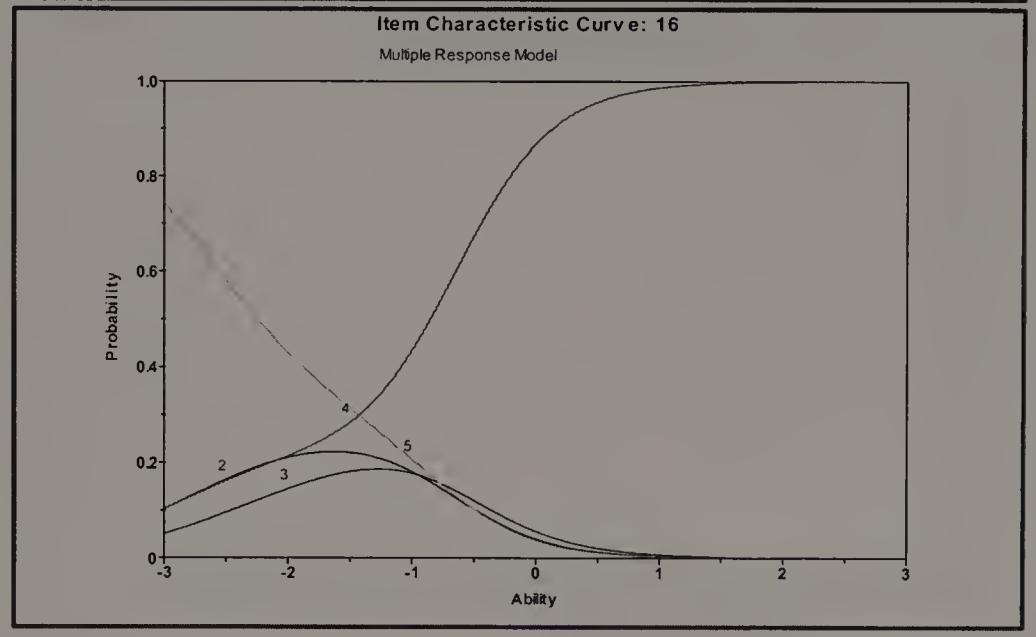
APPENDIX C

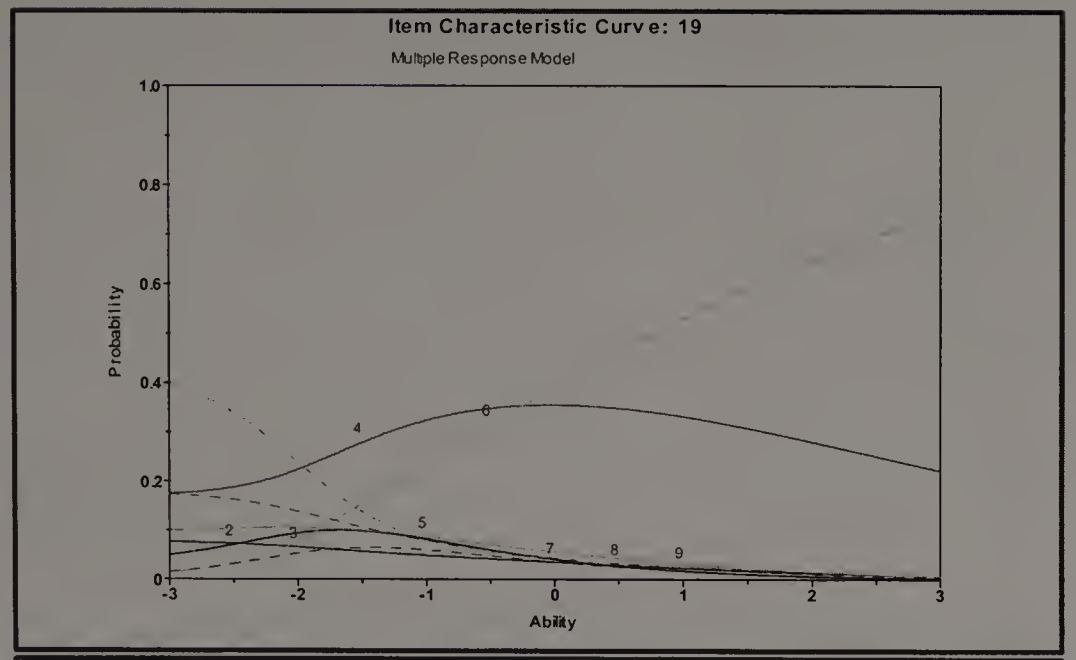
FORM A POPULATION ITEM CATEGORY RESPONSE FUNCTIONS (MULTIPLE-CHOICE MODEL)

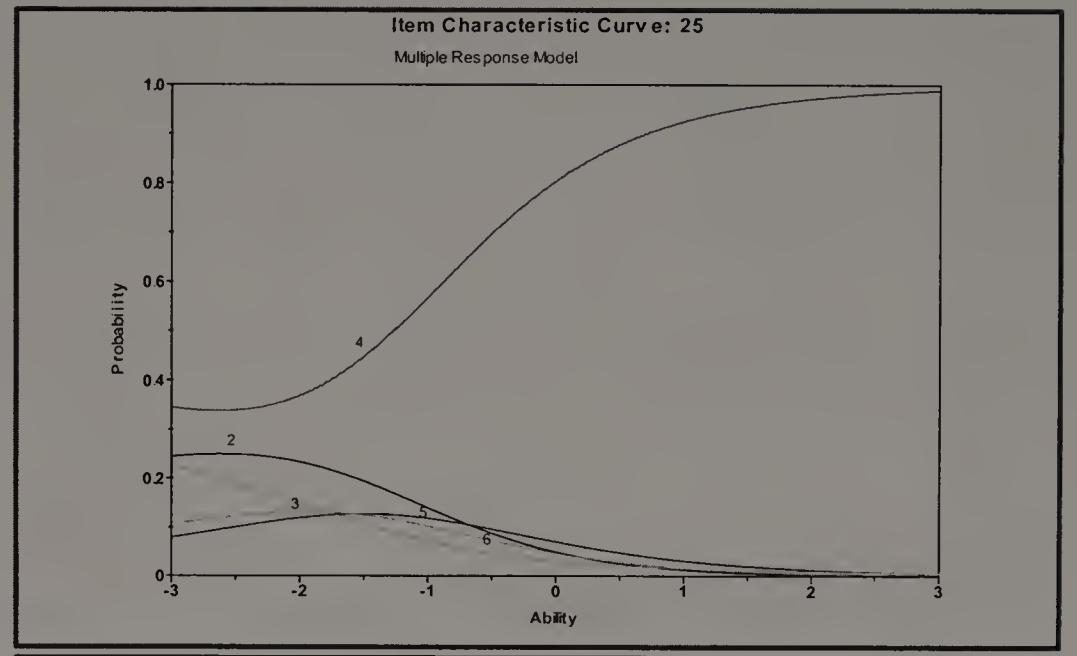

Item Characteristic Curve: 5
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6
7

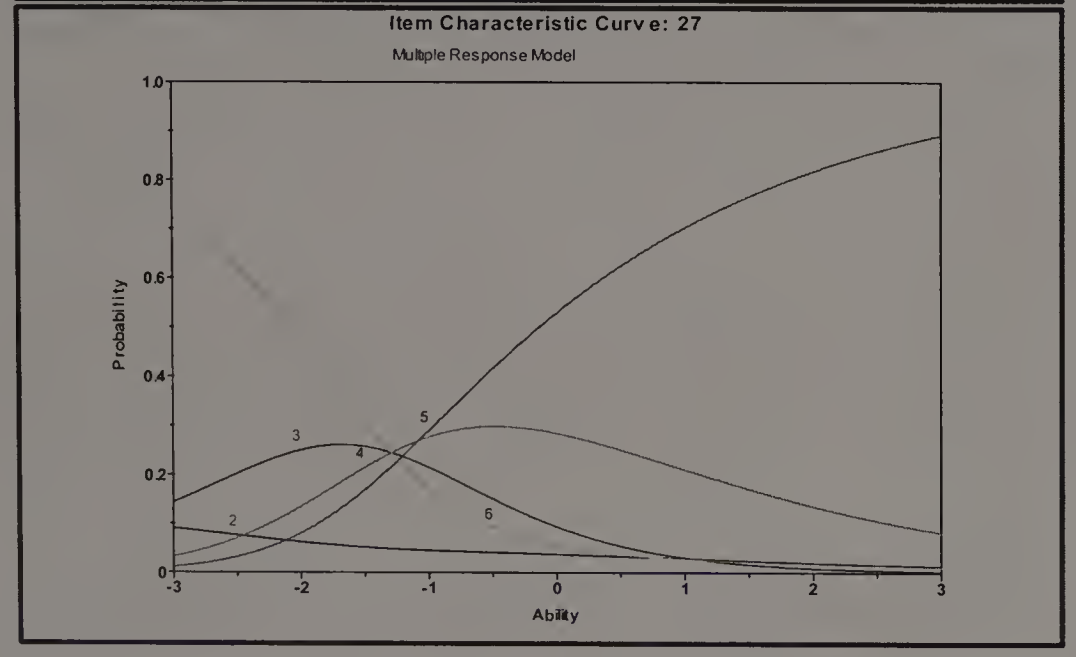

Item Characteristic Curve: 6
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6

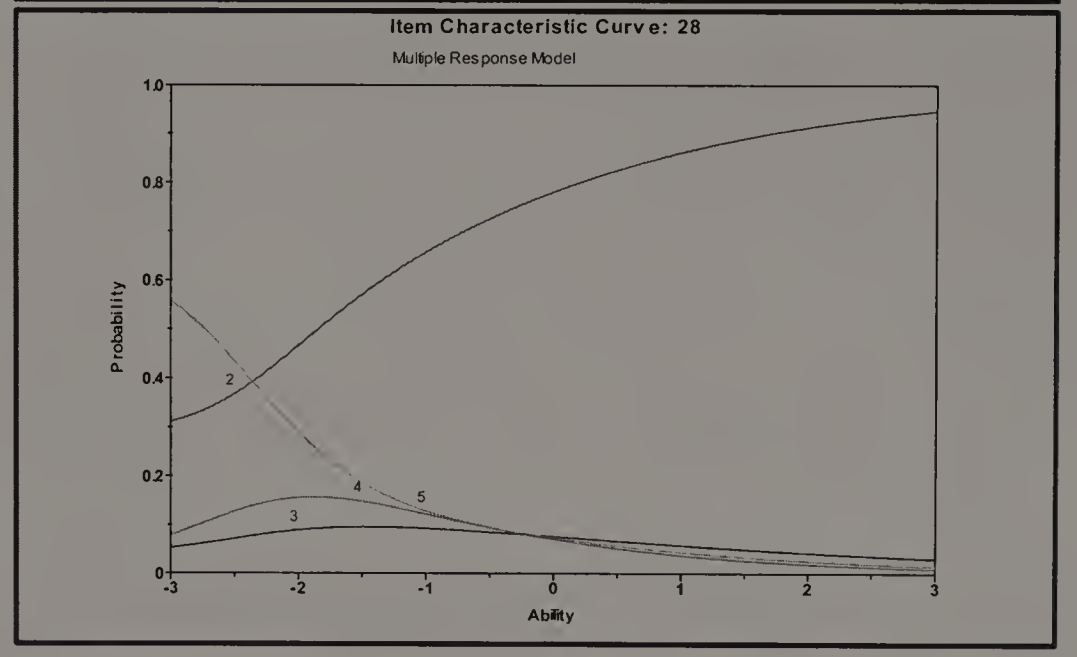

Item Characteristic Curve: 11
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3

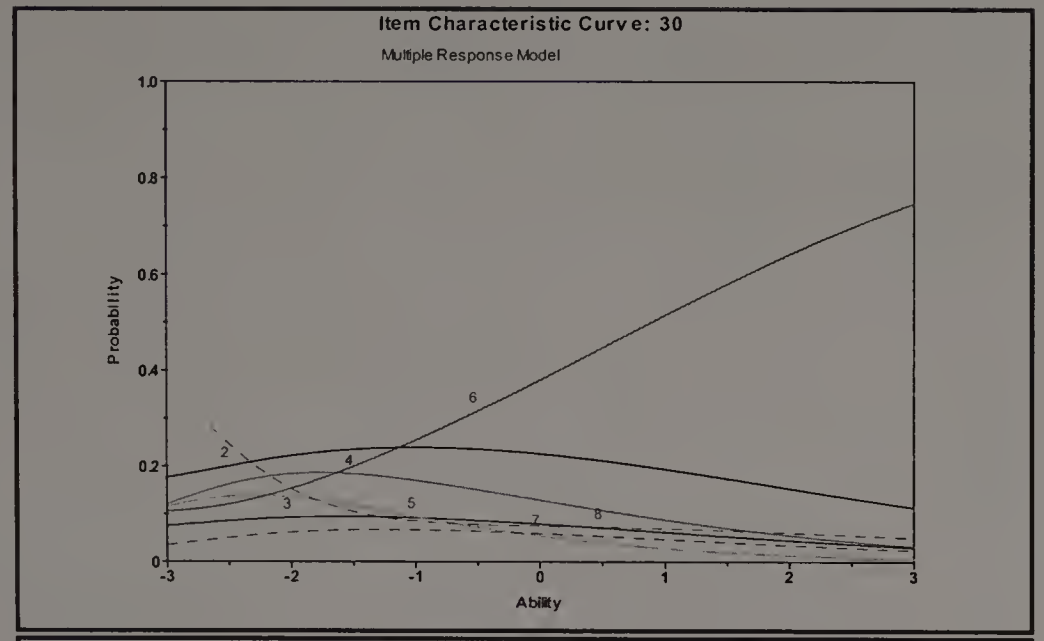

Item Characteristic Curve: 16
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5

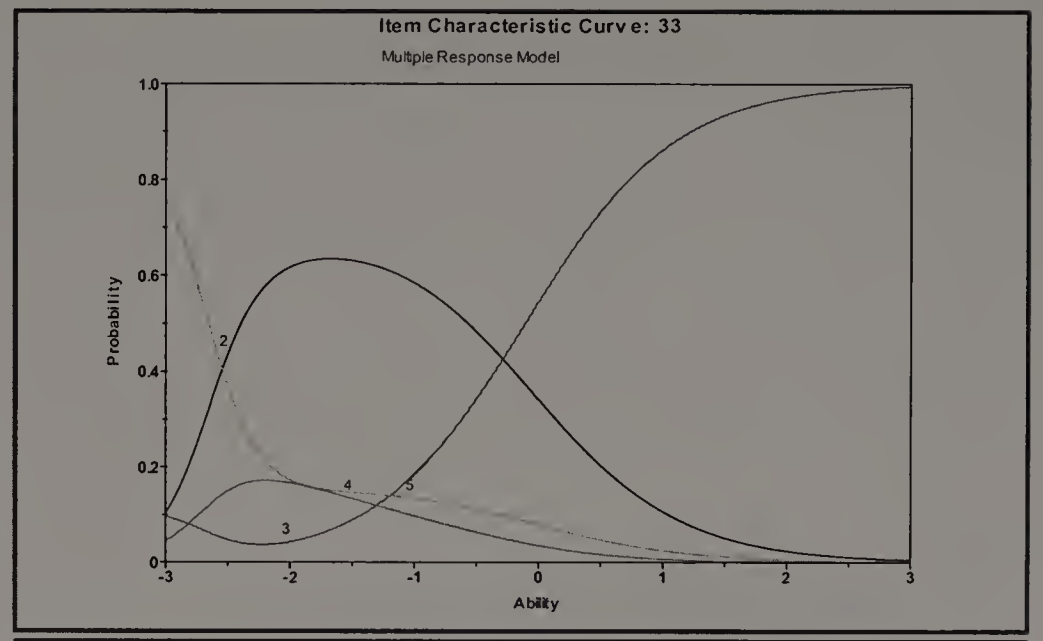

Item Characteristic Curve: 19
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6
7
8
9

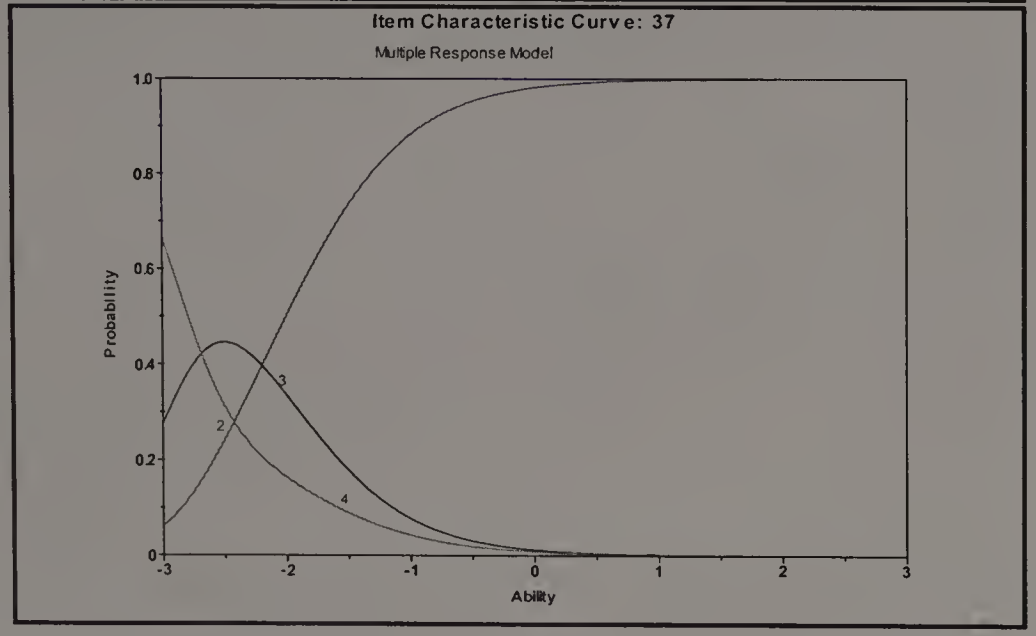

Item Characteristic Curve: 25
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6

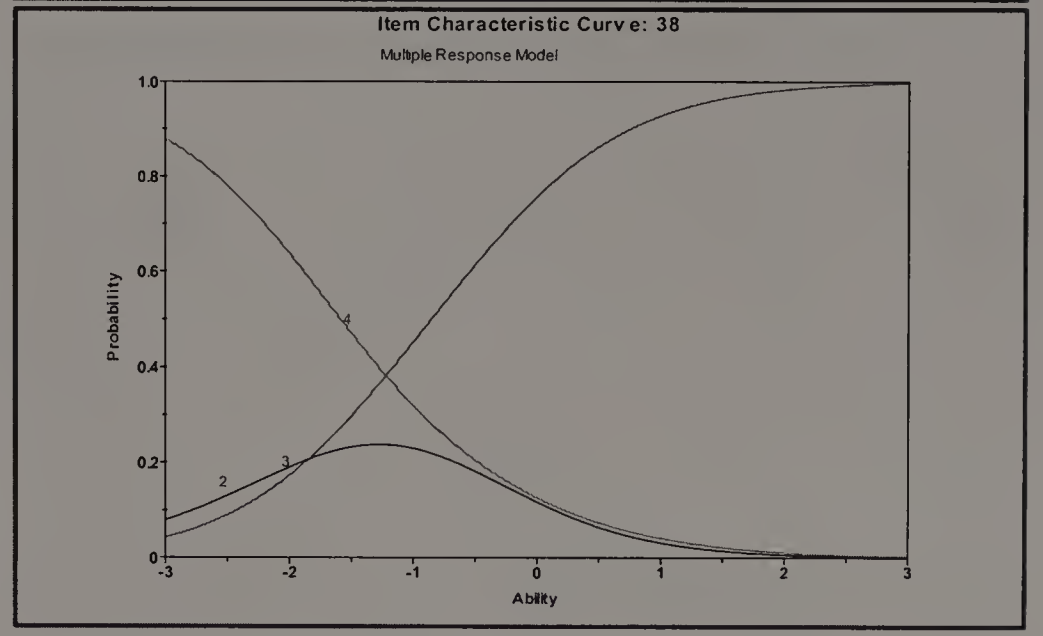

Item Characteristic Curve: 27
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6

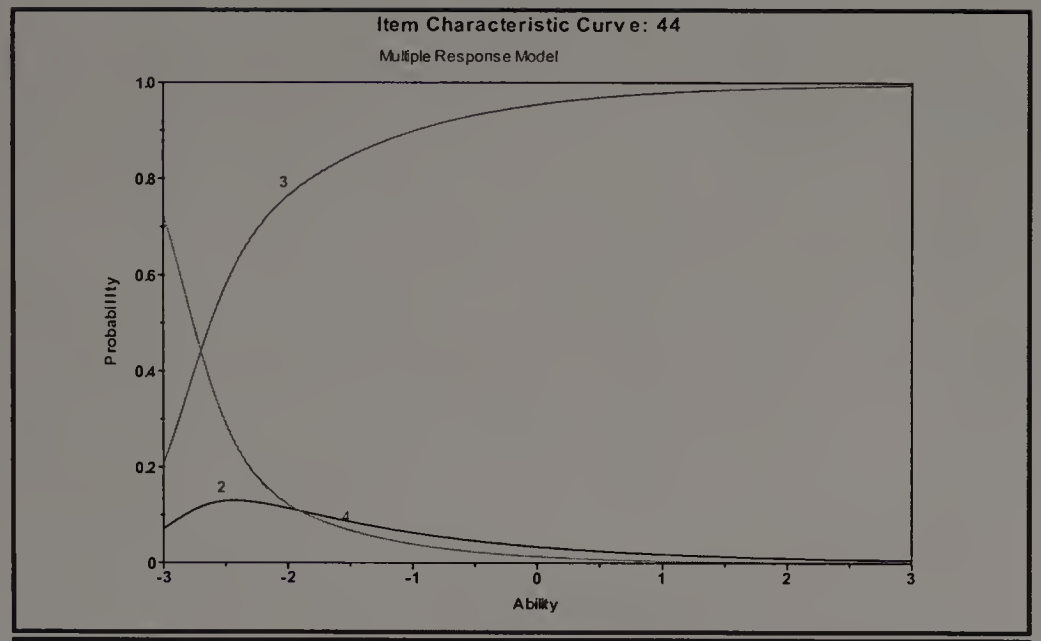

Item Characteristic Curve: 28
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5

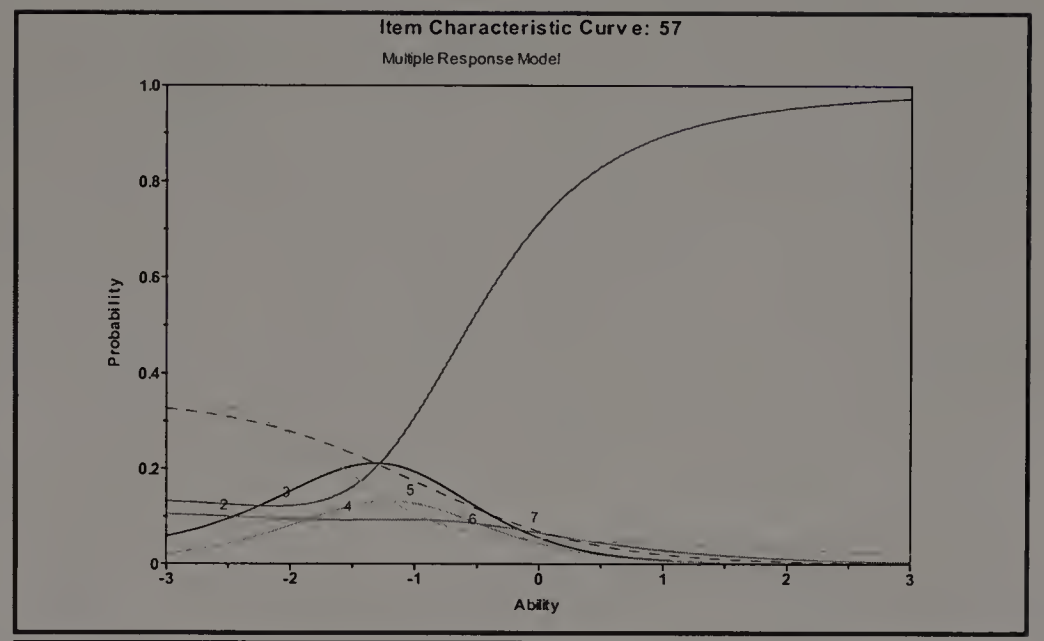

Item Characteristic Curve: 30
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6
7
8

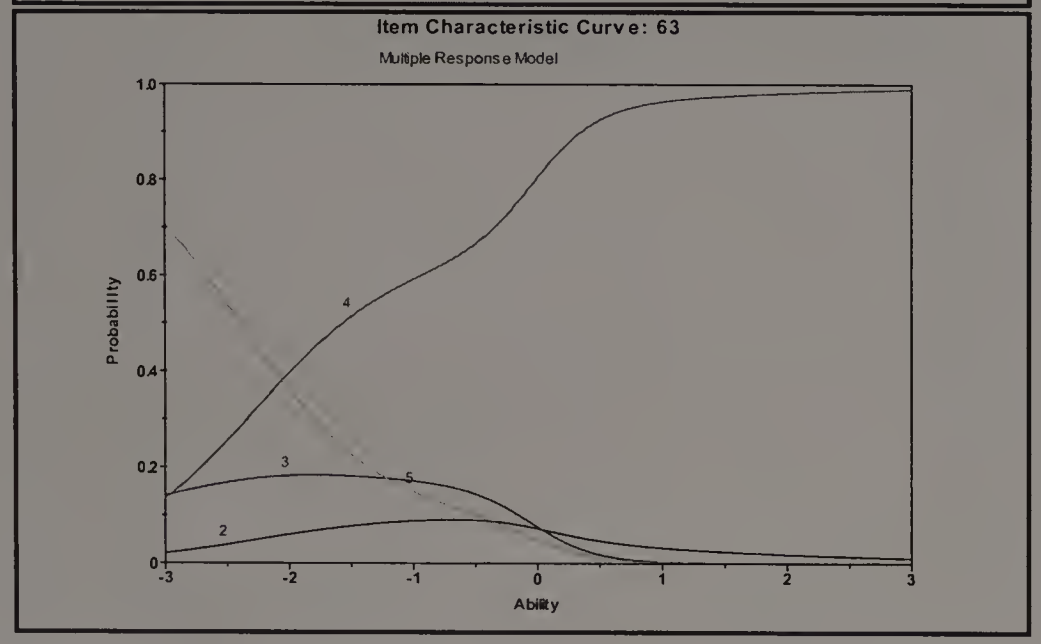

Item Characteristic Curve: 33
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5

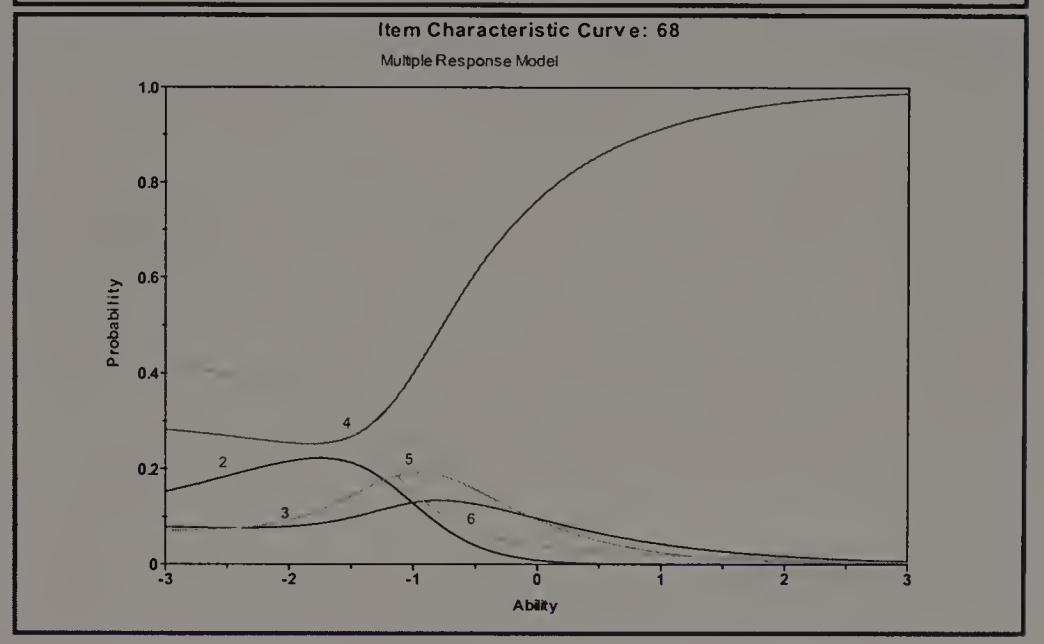

Item Characteristic Curve: 37
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4

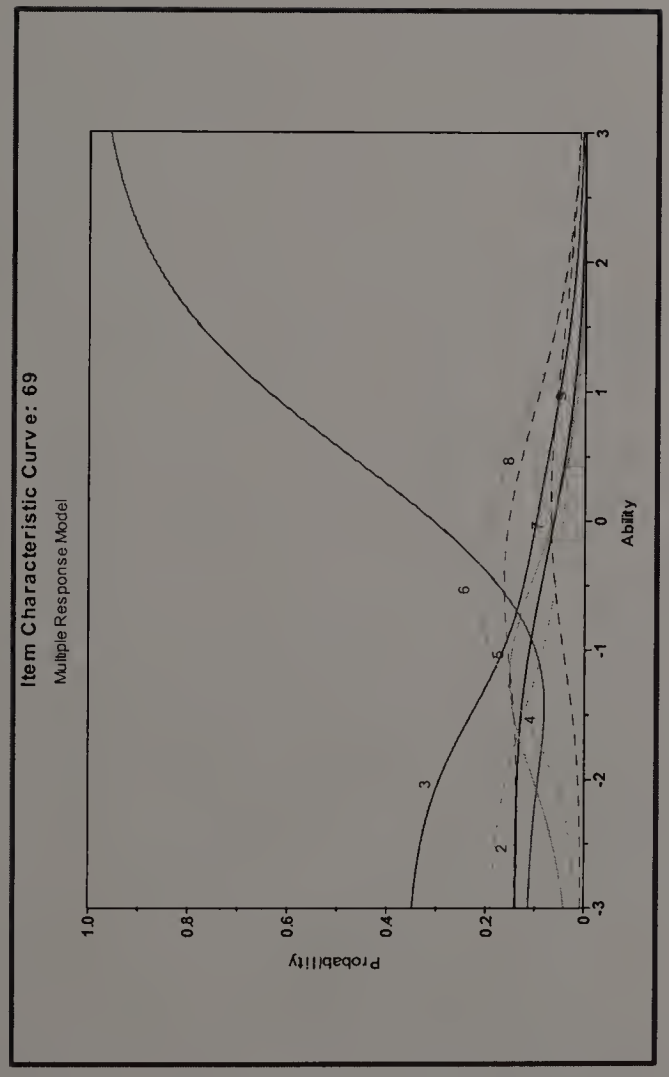

Item Characteristic Curve: 38
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4


Item Characteristic Curve: 44
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4


Item Characteristic Curve: 57
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6
7

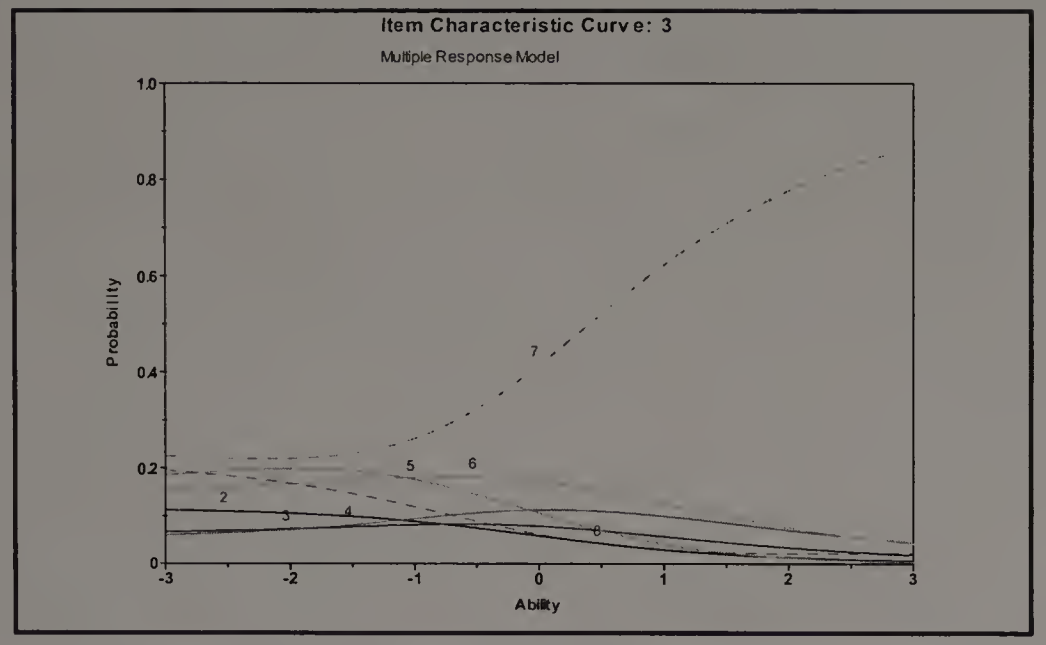

Item Characteristic Curve: 63
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
4
3
5
2

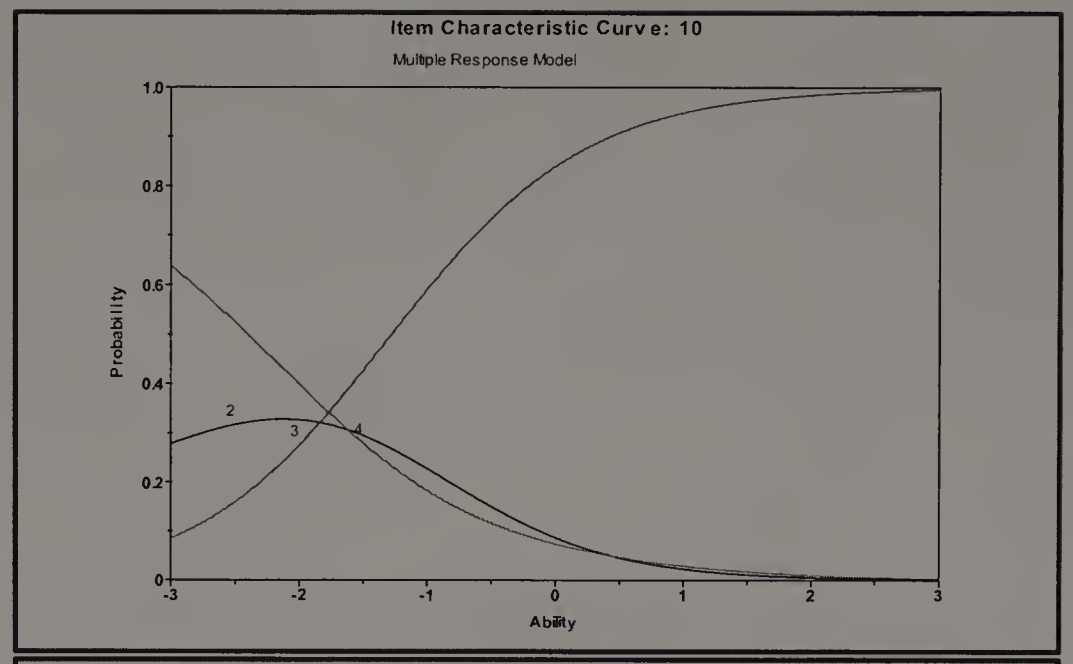

Item Characteristic Curve: 68
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
4
2
5
3
6

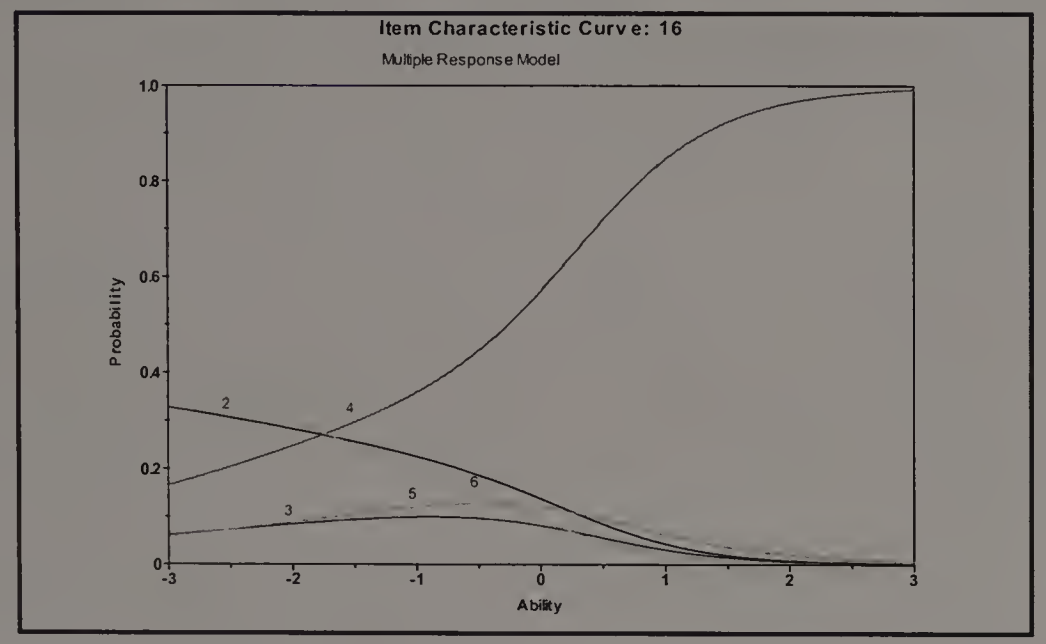

Item Characteristic Curve: 69
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6
8

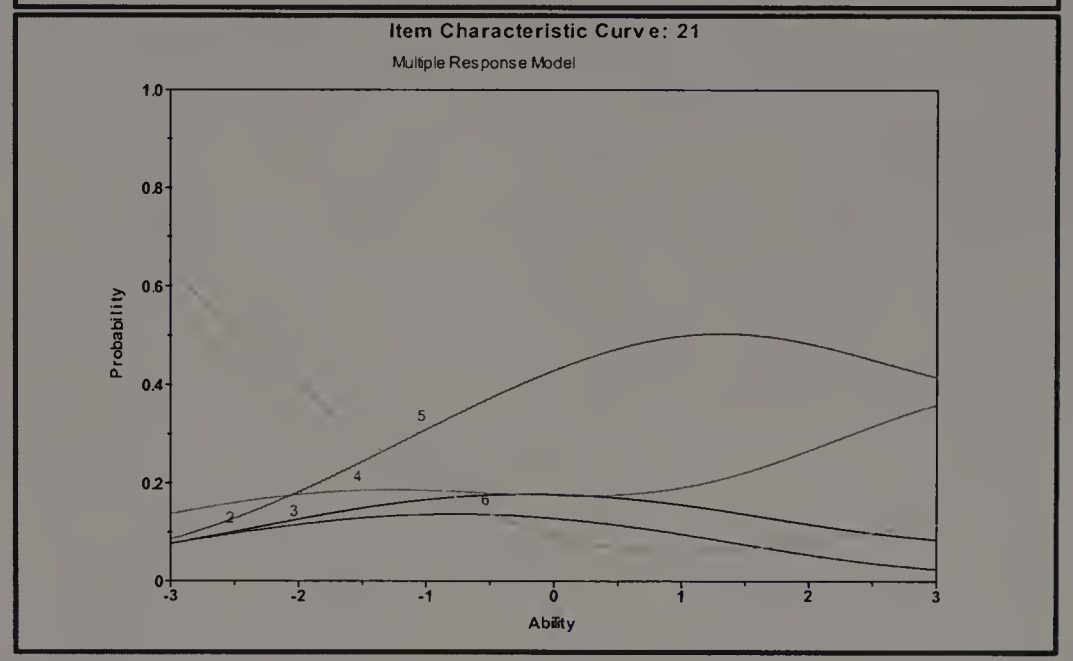
APPENDIX D

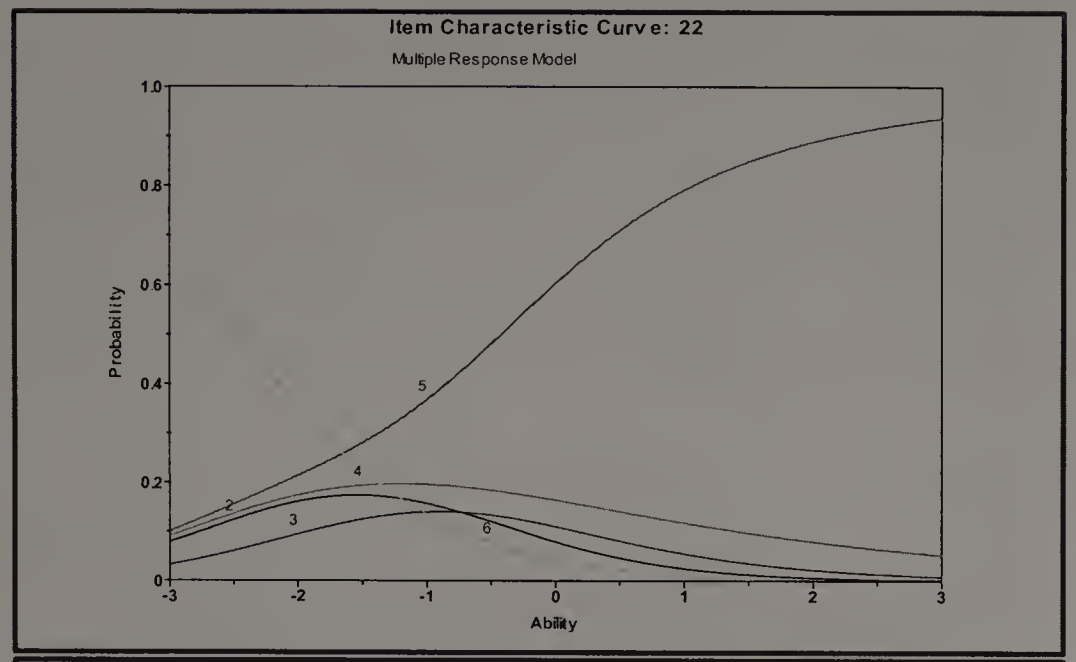
FORM F POPULATION ITEM CATEGORY RESPONSE FUNCTIONS (MULTIPLE CHOICE MODEL)

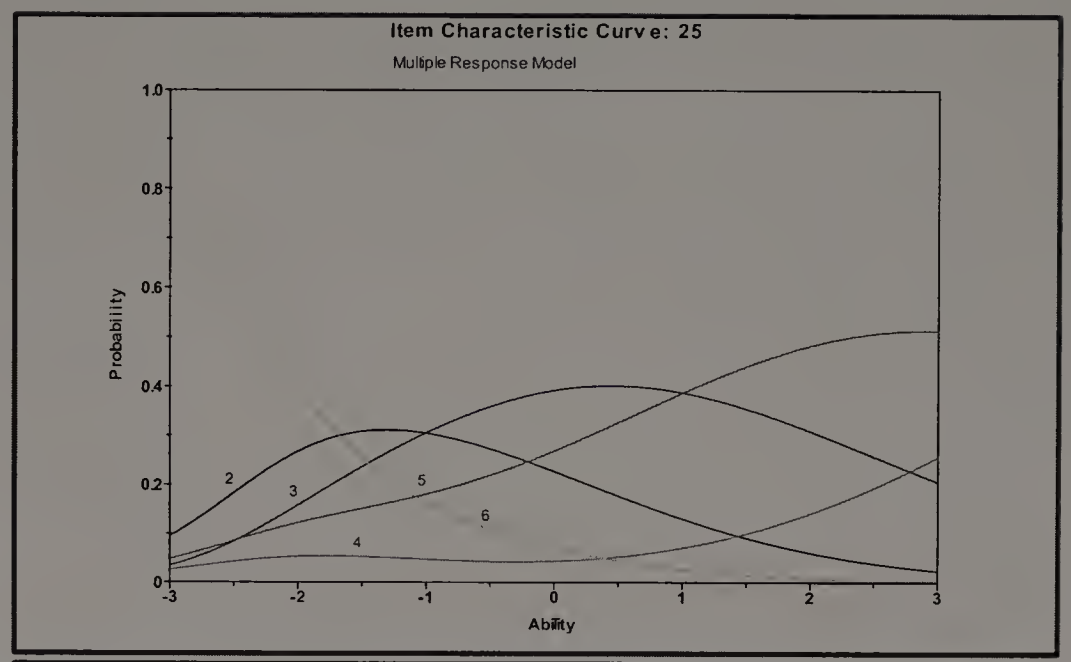

Item Characteristic Curve: 3
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6
7
8

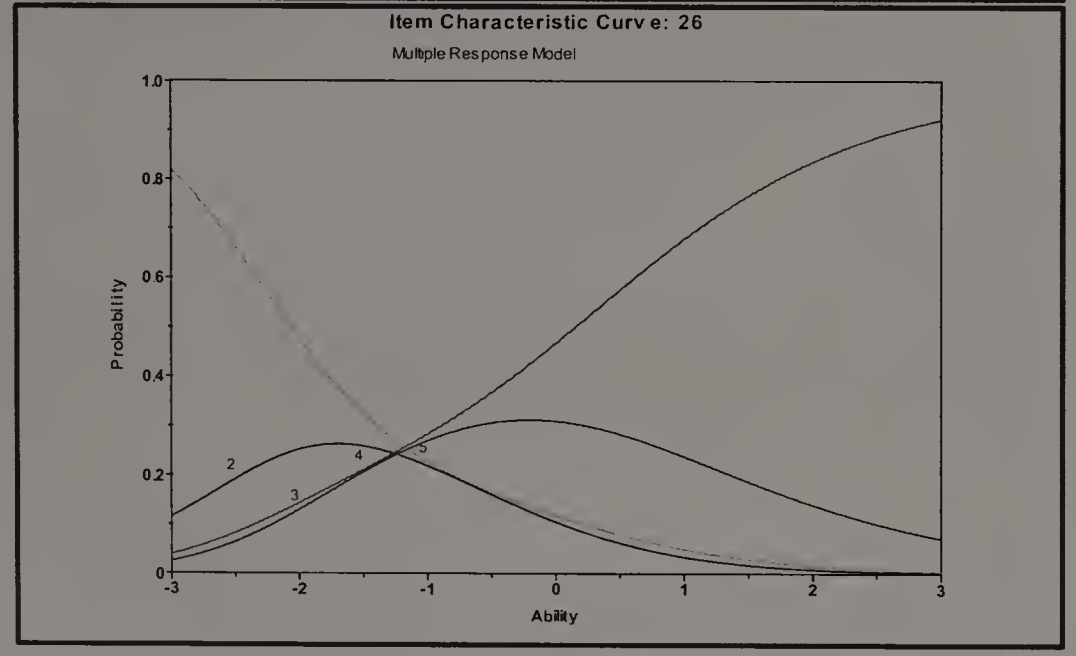

Item Characteristic Curve: 10
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4

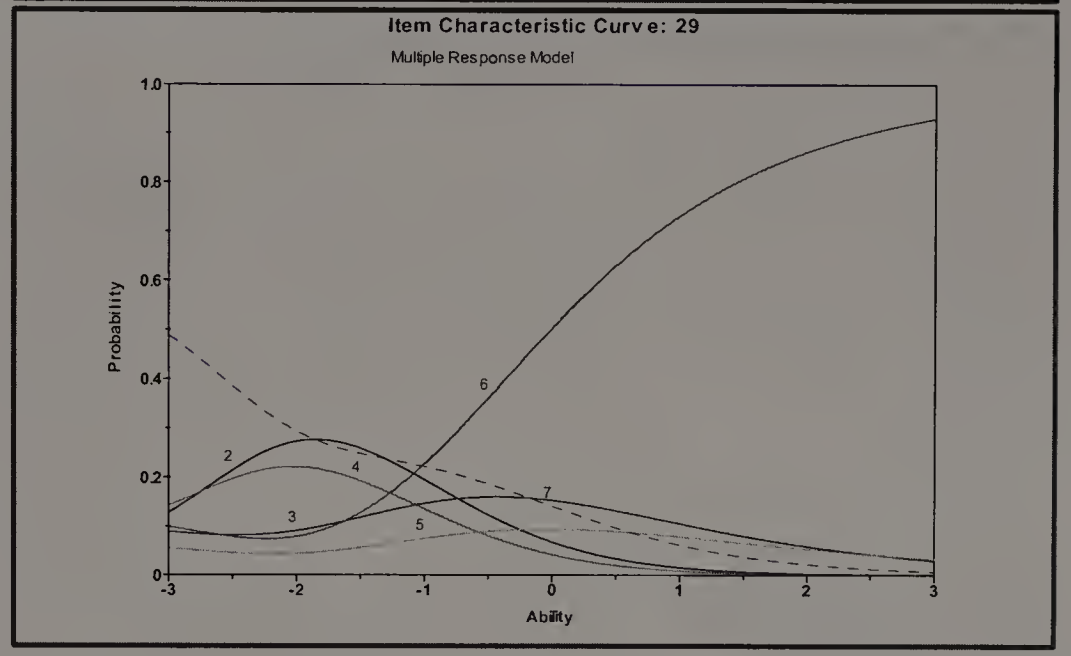

Item Characteristic Curve: 16
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6

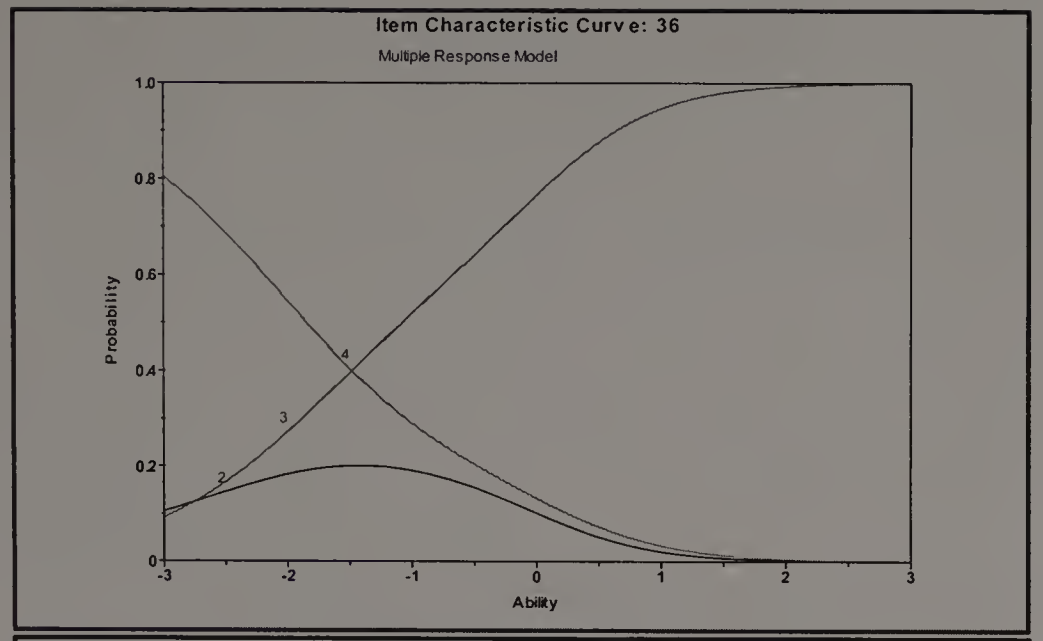

Item Characteristic Curve: 21
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6

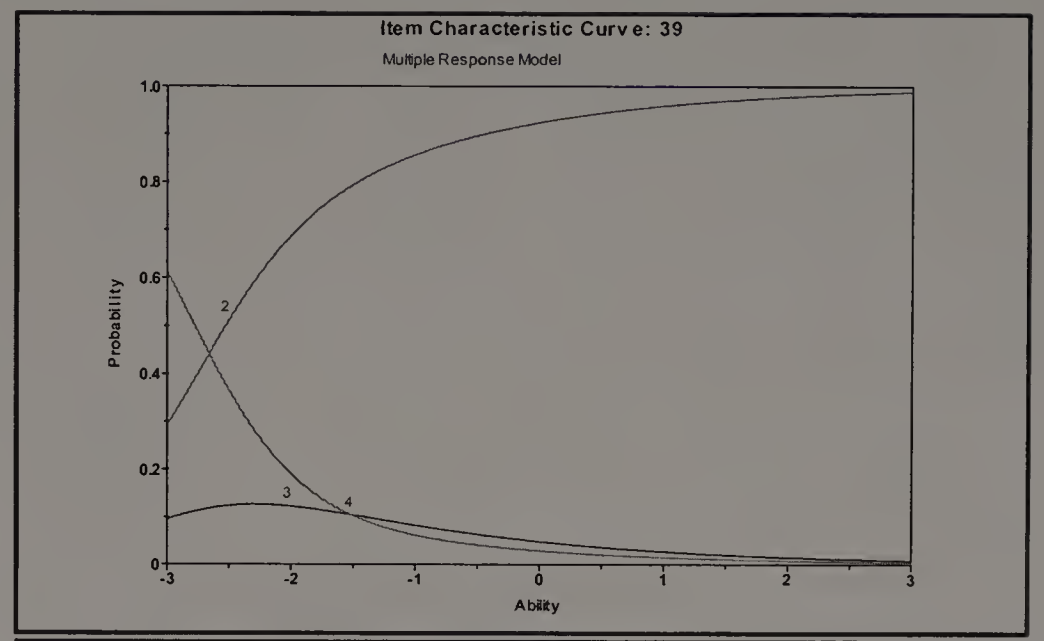

Item Characteristic Curve: 22
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6

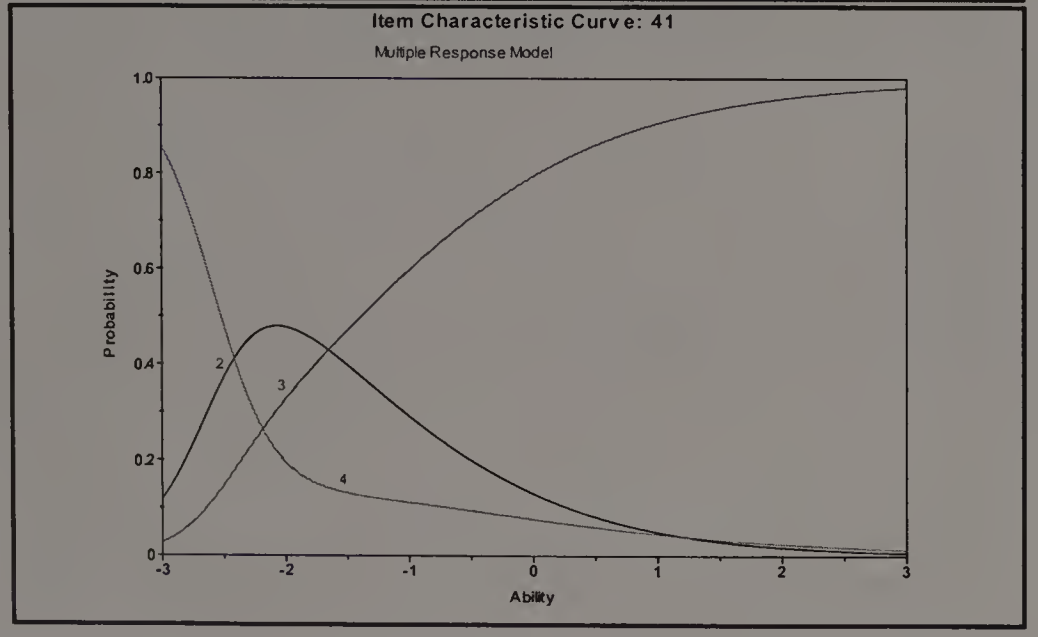

Item Characteristic Curve: 25
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6

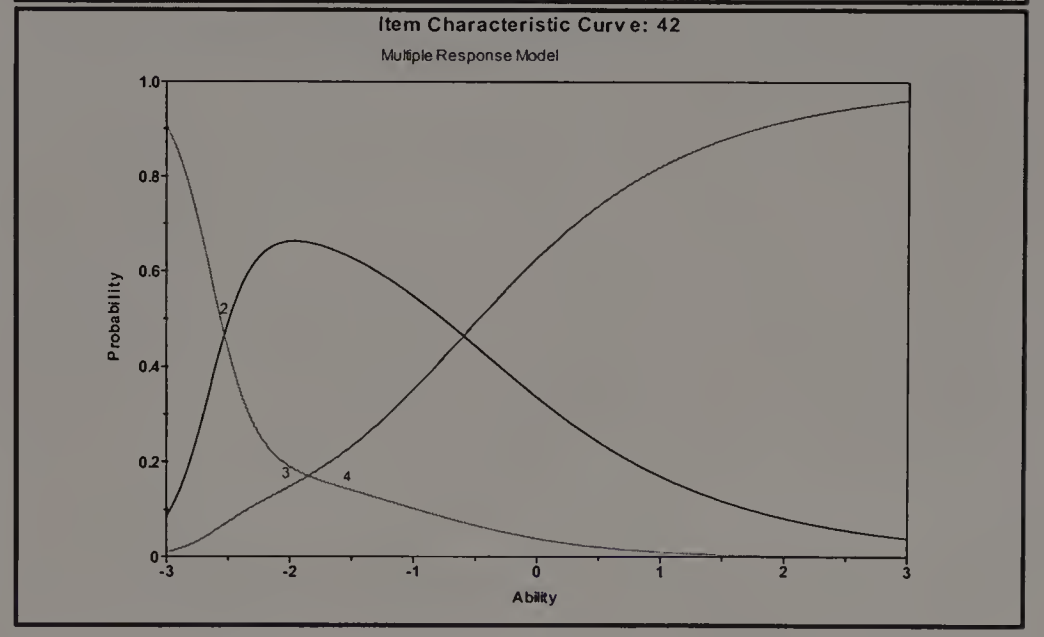

Item Characteristic Curve: 26
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5

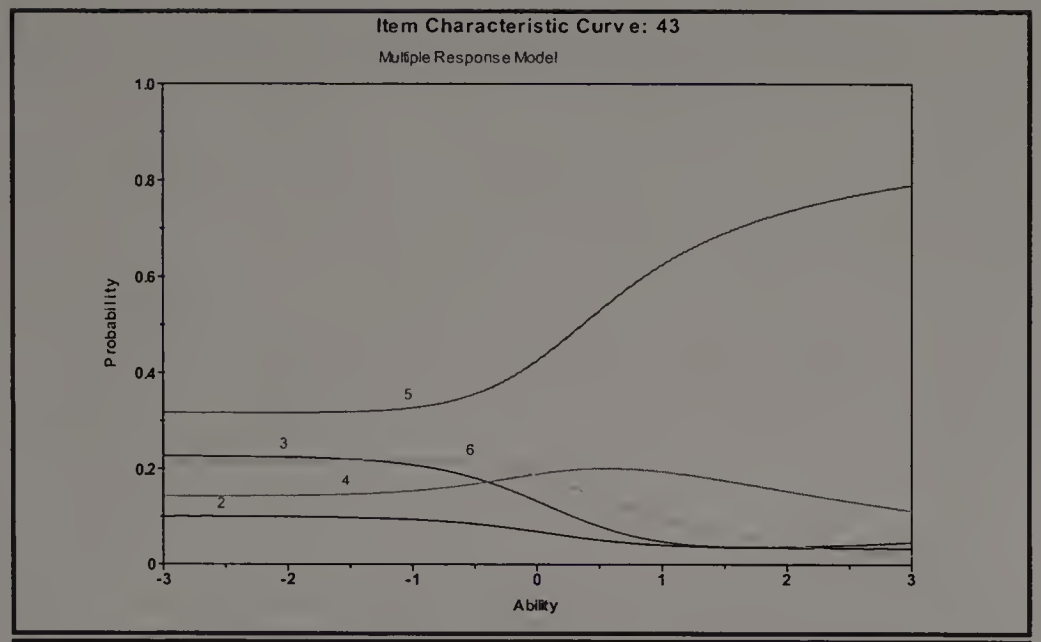

Item Characteristic Curve: 29
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6
7

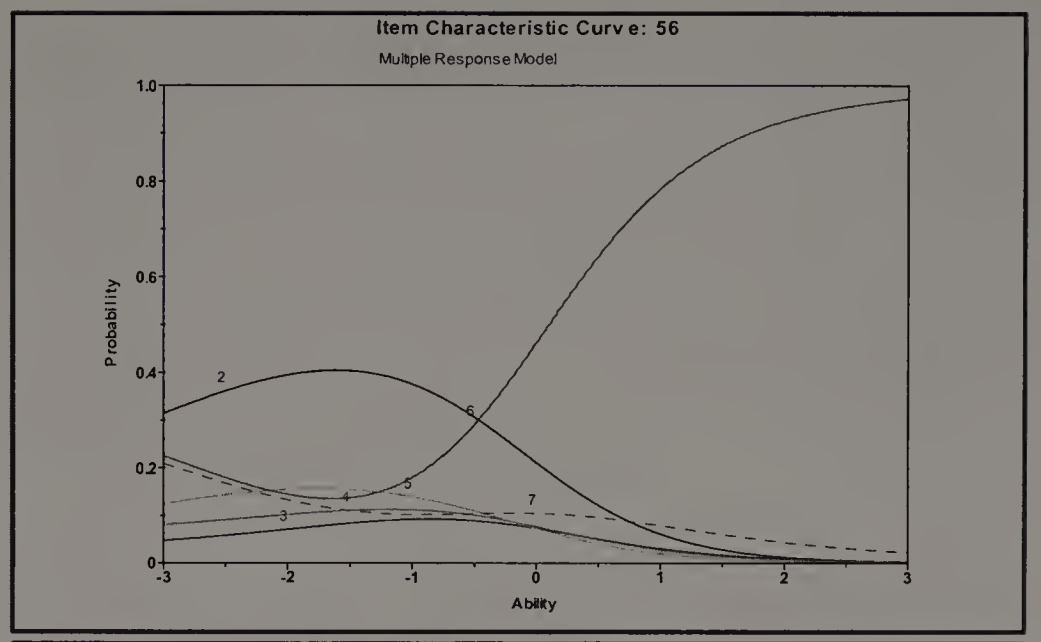

Item Characteristic Curve: 36
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4

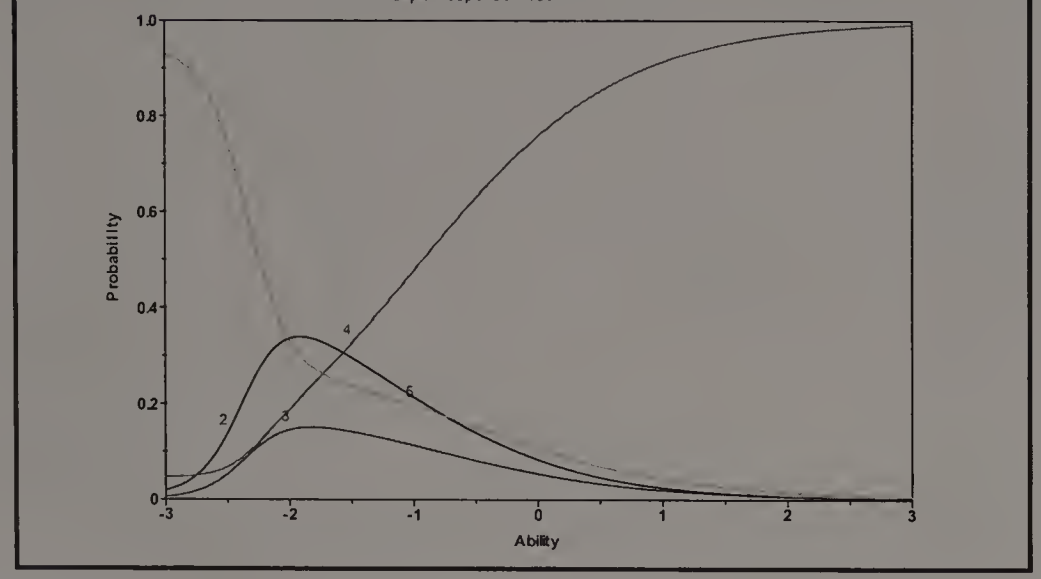

Item Characteristic Curve: 39
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4

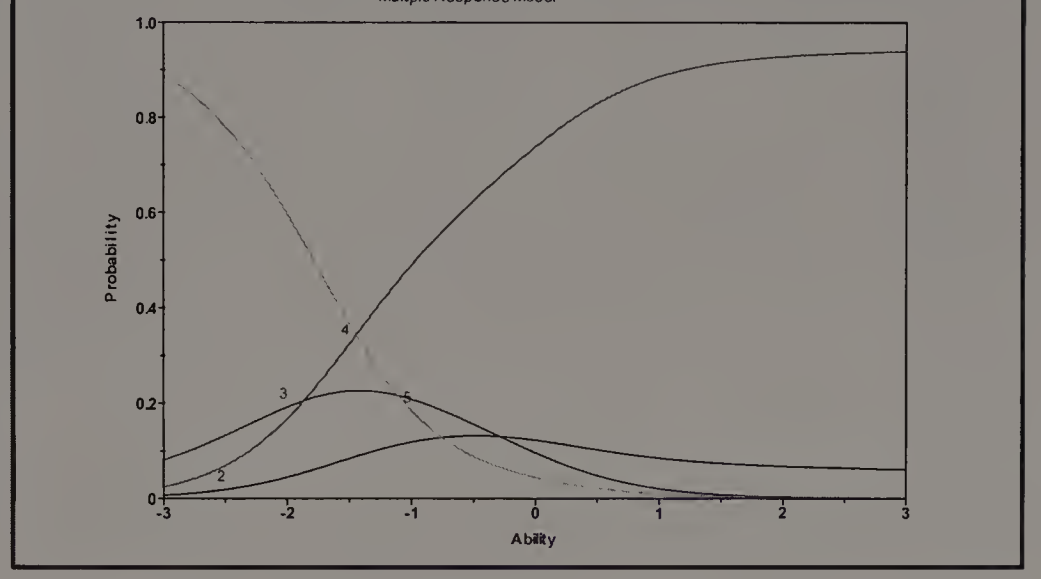

Item Characteristic Curve: 41
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4

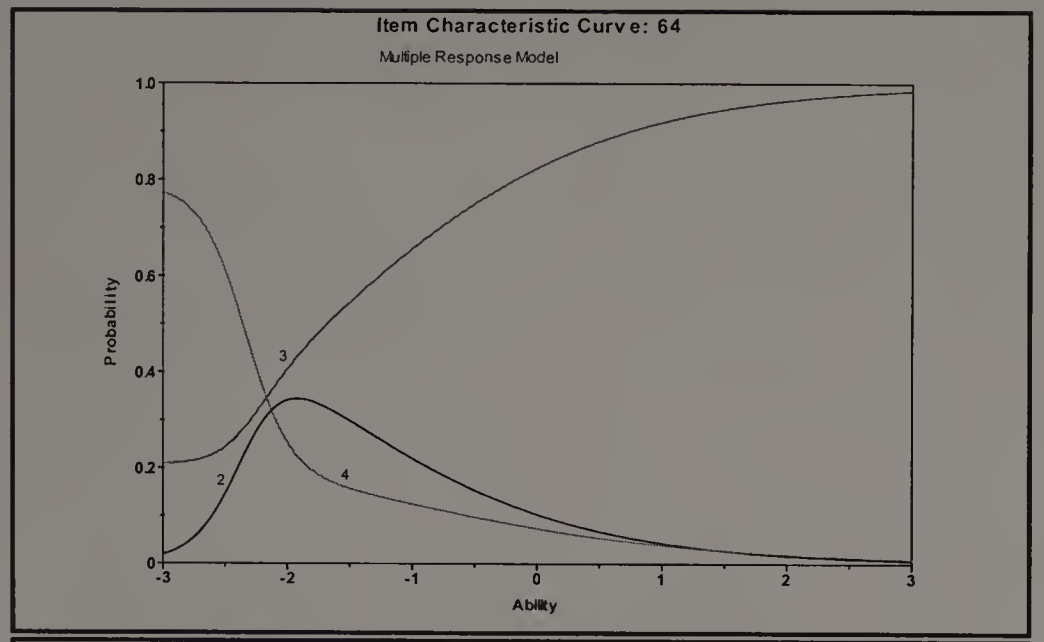

Item Characteristic Curve: 42
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4

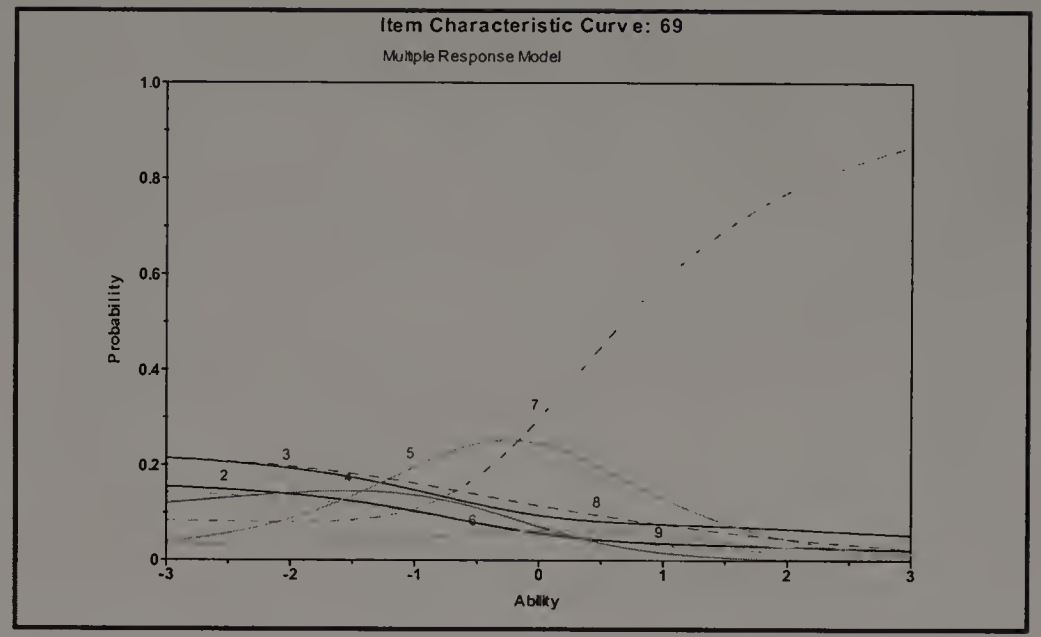

Item Characteristic Curve: 43
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
5
3
6
4
2

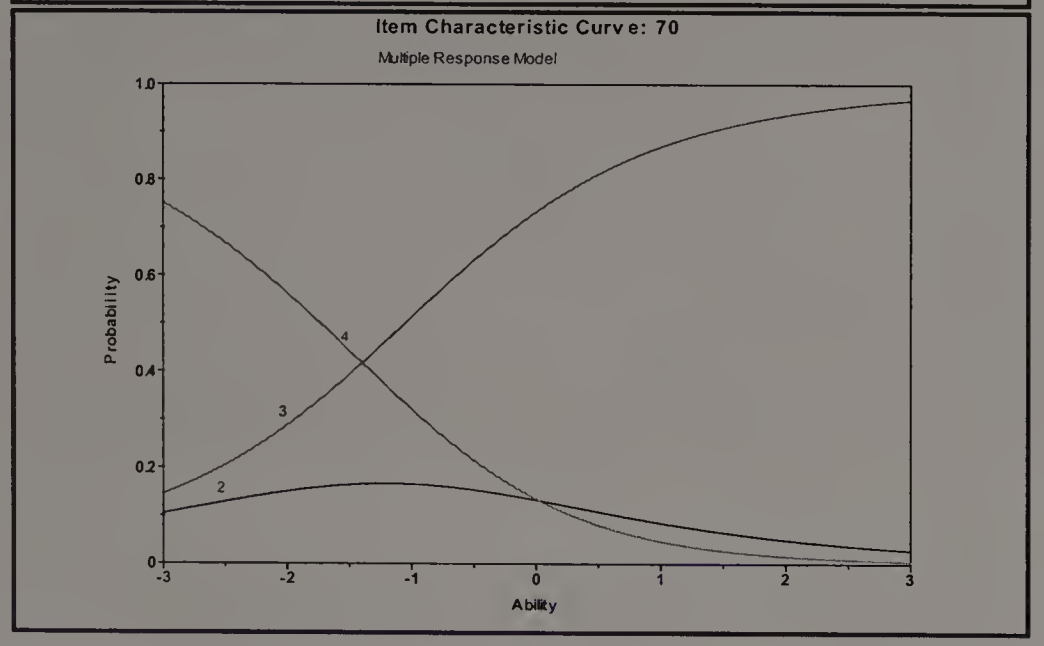

Item Characteristic Curve: 56
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
6
4
5
7
3


Item Characteristic Curve: 58
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
4
2
6
3


Item Characteristic Curve: 61
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
4
3
5
2


Item Characteristic Curve: 64
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4


Item Characteristic Curve: 69
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4
5
6
7
8
9


Item Characteristic Curve: 70
Multiple Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
2
3
4

BIBLIOGRAPHY

Albanese, M. A. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. Educational Measurement: Issues and Practices, 12(1), 28-33.

Baker, F. B. (1992). Item response theory: Parameter estimation techniques. New York: Marcel-Dekker.

Bejar, I., & Weiss, D. J. (1977). A comparison of empirical and differential option weighting scoring procedure as a function of inter-item correlation. Educational and Psychological Measurement, 37, 335-340.

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. Applied Psychological Measurement, 21(1), 65-88.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (chapters 17-20). Reading, MA: Addison-Wesley.

Blankenship, M. H., Cesare, S. J., Sympson, J. B. (1992, August). Test scoring in personnel selection: Number-correct scores vs. "polyscores." Paper presented at the 100th annual convention of the American Psychological Association, Washington, D. C.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Brown, W., & Thomson, G. H. (1925). The essentials of mental measurement (3rd ed.) London: Cambridge University Press.

Budescu, D. V., & Bar-Hillel, M. (1993). To guess or not to guess: A decision theoretic analysis of formula scoring. Journal of Educational Measurement, 30, 277-291.

Claudy, J. G. (1978). Biserial weights: A new approach to test item option weighting. Applied Psychological Measurement, 2, 25-30.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), Handbook of clinical psychology. New York: McGraw-Hill.

Coombs, C. H. (1953). On the use of objective examinations. Educational and Psychological Measurement, 13, 308-310.

Coombs, C. H., Milholland, J. E. & Womer, J. F. B. (1956) The assessment of partial knowledge. Educational and Psychological Measurement, 16, 13-37.

Crocker, L., Algina, J. (1986). Introduction to classical and modern test theory. New York: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1941). An experimental comparison of the multiple true-false and multiple-multiple-choice test. Journal of Educational Psychology, 32, 533-543.

Cross, L. H., Ross, F. K., & Geller, E. S. (1980). Using choice-weighted scoring of multiple-choice tests for determination of grades in college courses. Journal of Experimental Education, 48, 296-301.

Davey, T., Godwin, J., & Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. Journal of Educational Measurement, 34(1), 21-41.

Davis, F. B., & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 19, 159-170.

De Ayala, R. J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. Educational and Psychological Measurement, 49, 789-805.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. Applied Psychological Measurement, 16, 327-343.

DeFinetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 18, 87-123.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. Journal of Educational Measurement, 31, 295-311.

Downey, R. G. (1979). Item-option weighting of achievement tests: Comparative study of methods. Applied Psychological Measurement, 3, 453-461.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. Applied Psychological Measurement, 19, 143-165.

Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E., & Candell, G. (1989). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. Applied Psychological Measurement, 13, 285-299.

Dressel, P. L., & Schmid, J. (1953). Some modifications of multiple-choice items. Educational and Psychological Measurement, 13, 574-595.

Ebel, R. L. (1965). Confidence weighting and test reliability. Journal of Educational Measurement, 2, 49-57.

Ebel, R. L. (1978). The ineffectiveness of multiple true-false test items. Educational and Psychological Measurement, 38, 37-44.

Echternacht, G. J. (1972). The use of confidence testing in objective testing. Review of Educational Research, 42, 217-236.

Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. Applied Measurement in Education, 2, 79-96.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. Educational Measurement: Issues and Practices, 11(4), 21-26.

Gage, N. L. (1957). Logical vs. empirical scoring keys: The case of the MTA1. Journal of Educational Psychology, 48, 213-216.

Glasnapp, D. G., & Poggio, J. P. (1994, April). Psychometric characteristics of the multiple-correct multiple-choice item. Paper presented at annual meeting of the National Council on Measurement in Education, New Orleans.

Grosse, M. E., & Wright, B. D. (1985). Validity and reliability of true-false tests. Educational and Psychological Measurement, 45, 1-13.

Guilford, J. P. (1941). A simple scoring weight for test items and its reliability. Psychometrika, 2, 15-21.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.) Prediction of Personal Adjustment. Social Science Research Bulletin, 48, 321-345.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 2, 51-78.

Haladyna, T. M., & Sympson, J. B. (1988, April). Empirically based polychotomous scoring of multiple-choice test items: A review. In C. E. Davis (Chair), New developments in polychotomous scoring and modeling. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 147-200). New York: Macmillan.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. Journal of Educational Measurement, 7, 75-82.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hanna, G. S. (1975). Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. Journal of Educational Measurement, 12, 175-178.

Hendrickson, G. F. (1971). The effect of differential option weighting on multiple-choice objective tests. Journal of Educational Measurement, 8, 291-296.

Holmes, P. (2002). Multiple evaluation versus multiple-choice as testing paradigm – feasibility, reliability and validity in practice. Unpublished doctoral dissertation, University of Twente, the Netherlands.

Hsu, T. C., Moss, P. A., & Khampalikit, C. (1984). The merits of multiple-answer item as evaluated by using six scoring formulas. Journal of Experimental Education, 52, 152-158.

Hubbard, J. P. (1978). Measuring medical education: The tests and the experience of the National Board of Medical Examiners (2nd edition). Philadelphia: Lea & Febiger.

Hutchinson, T. P. (1982). Some theories of performance in multiple-choice tests, and their implications for variants of the task. British Journal of Mathematical and Statistical Psychology, 35, 71-89.

Huynh, H., & Casteel, J. (1987). The usefulness of the Bock model for scoring with information from incorrect responses. Journal of Experimental Education, 55, 131-136.

Jacob, P. I., & Vanderventer, M. (1968). Information in wrong responses (Research Bulletin, 68-25), Princeton, N. J.: Educational Testing Service.

Jaradat, D., & Tollefson, N. (1988). The impact of alternative scoring procedures for multiple-choice items on test reliability, validity and grading. Educational and Psychological Measurement, 48, 627-635.

Jodoin, M. G. Measurement efficiency of innovative item formats in computer-based testing. Journal of Educational Measurement, 40, 1-15.

Kansup, W., & Hakstian, A. R. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. Journal of Educational Measurement, 12, 219-239.

Keller, L. A., Swaminathan, H., & Sireci, S. G. Evaluating scoring procedures for context-dependent item sets. Applied Measurement in Education, 16, 207-222.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. Educational and Psychological Measurement, 43, 675-685.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mislevy, R. J. (1995). Foundations of a new test theory. In N. Fredericksen., R. J. Mislevy, & I. I. Bejar (Eds.), Test theory for a new generation of tests (pp. 19-39). Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J., & Bock, R. D. (1991). BILOG 3.5. Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Nedelsky, L. (1954). Ability to avoid gross error as a measure of achievement. Educational and Psychological Measurement, 14, 459-472.

Nishisato, S. (1980). Analysis of categorical data: Dual scaling and its application. Toronto, Canada: University of Toronto Press.

O'Neill, K., & Folk, V. (1996, April). Innovative CBT item formats in a teacher licensing program. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Orleans, J. S., & Sealy, G. A. (1928). Objective tests. Yonkers, New York: World Book Company.

Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. Academic Medicine, 70(3), 194-201.

Parshall, C. G., Davey, T., & Pashley, P. (2001). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds), Computerized adaptive testing: Theory and practice (pp. 129-148). Boston, MA: Kluwer Academic Publishers.

Parshall, C. G., Stewart, R., & Ritter, J. (1996, April). Innovations: Sound, graphics, and alternative response modes. Paper presented at the annual meeting of the National Council on Measurement in Education. New York.

Patnaik, D., & Traub, R. E. (1973). Differential weighting by judged degree of correctness. Journal of Educational Measurement, 10, 281-286.

Pomplun, M., & Omar, MD. H. (1997). Multiple-mark items: An alternative objective item format? Educational and Psychological Measurement, 57(6), 949-962.

Raffeld, P. (1975). The effects of Guttman weights on the reliability and predictive validity of objective tests when omissions are not differentially weighted. Journal of Educational Measurement, 12, 179-185.

Reckase, M. D. (1995). The reliability of ratings versus the reliability of scores. Educational Measurement: Issues and Practice, 14, 31.

Reilly, R. R., & Jackson, R. (1973). Effects of empirical option weighting on validity and reliability on an academic aptitude test. Journal of Educational Measurement, 10, 185-194.

Sabers, D. L., & White, G. W. (1969). The effects of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. Journal of Educational Measurement, 6, 93-96.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika (Monograph Supp. No. 17).

Samejima, F. (1976). Graded response model of the latent trait theory and tailored testing, In C. K. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing (pp. 5-17). Washington, DC: U.S. Government Printing Office.

Samejima, F. (1979). A new family of models for the multiple-choice item (Research Rep. No.79-4). University of Tennessee, Department of Psychology.

Serlin, R. C., & Kaiser, H. F. (1978). A method for increasing the reliability of a short multiple-choice test. Educational and Psychological Measurement, 38, 337-340.

Shuford, E. H., Jr., Albert, A., & Massengill, H. E. (1966). Admissable probability measurement procedures. Psychometrika, 31, 125- 145.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28(3), 237-247.

Smith, R. M. (1970). Assessing partial knowledge in vocabulary. Journal of Educational Measurement, 24, 217-231.

Sympson, J. B. (1983, June). A new IRT model for calibrating multiple-choice items. Paper presented at the annual meeting of the Psychometric Society, Los Angeles, CA.

Sympson, J. B. (1988, May). A procedure for linear polychotomous scoring of test items. Paper presented at the Office of Naval Research Contractors' Meeting on Model-based Psychological Measurement. Iowa City, Iowa.

Sympson, J. B. (1990). POLY: A computer program for polychotomous item analysis (Release 06/15/90). San Diego: Navy Personnel Research and Development Center.

Sympson, J. B., & Davison, M. L. (1989, March). Reducing test length with polychotomous scoring. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Sympson, J. B., & Haladyna, T. M. (1988, April). An evaluation of "polyweighting" in domain-referenced testing. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement, 13, 201-214.

Thissen, D. M. (2002). MULTILOG 7.0. Multiple, categorical item analysis and test scoring using item response theory. Chicago, IL: Scientific Software.

Thissen, D. M., & Steinberg, L. (1984). A response model for multiple-choice items. Psychometrika, 49, 501-519.

Thissen, D. M., & Steinberg, L. (1986). Taxonomy of item response models. Psychometrika, 51, 567-577.

Thissen, D. M., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. Journal of Educational Measurement, 26, 161-176.

Thompson, T. D., Pommerich, M. (1996, April). Examining the sources and effects of local dependence. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometric for testlets. Journal of Educational Measurement, 27, 1-14.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. Review of Educational Research, 40, 663-705.

Wilcox, R. R. (1981). Solving measurement problems with an answer until correct scoring procedure. Applied Psychological Measurement, 5, 399-413.

Yamamoto, K., & Kulick, E. (1992, April). An information-based approach to maintaining content validity and determining the relative value of polytomous and dichotomous items. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Yee, A. H., & Kriewall, T. A. (1969). New logical scoring key for the Minnesota Teacher Attitude Inventory. Journal of Educational Measurement, 6, 11-14.

Yen, W. M. (1984). Effects of local item dependencies on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8(2), 125-145.

Yen, W. M. (1993). Scaling performance assessment: Strategies for managing local item dependence. Journal of Educational Measurement, 30(3), 187-214.